Statistical Methods in Clinical Trial Design

by

Yu Du

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

March, 2018

Abstract

Numerous human medical problems or diseases have been aided by the development of effective treatments such as drugs and medical devices. Clinical trials are an integral part of the development process, determining the safety and efficacy of the new proposed treatment, as required by the Food and Drug Administration of the United States. A reliable, efficient and cost-effective way of conducting the clinical trials is important for advancing useful treatments/devices to market and screening out the useless ones, thus benefiting public health in a timely manner. I developed several statistical methods and applications toward this purpose, ranging from early, small scale Phase I studies to late, large scale Phase III studies in clinical trials.

In Phase I studies, I establish a general framework for a multi-stage adaptive design where I jointly model a continuous efficacy outcome and continuous toxicity endpoints from multiple treatment cycles, unlike the traditional method that only considers a binary toxicity endpoint (joint work with Mayo Clinic). Extensive simulations confirmed that the design had a high probability of making the correct

dose selection and good overdose control. To our best knowledge, this proposed Phase I dual-endpoint dose-finding design is the first to incorporate multiple cycles of toxicities and a continuous efficacy outcome.

I also propose and evaluate a two-stage, adaptive clinical trial design for Phase II studies. Its goal is to determine whether future phase 3 (confirmatory) trials should be conducted, and if so, which population should be enrolled. I compute an approximate Bayes optimal design considering a combination of future health benefits and costs.

Turning to Phase III studies, I analyze the performance of adaptive enrichment designs with delayed outcome, leveraging information in baseline variables and short-term outcomes to improve precision by using semiparametric, locally efficient estimators at each interim analysis. I also propose a prediction method for analyzing heterogeneity in treatment response, as a secondary analysis, through the identification of treatment covariate interactions honoring different hierarchical conditions.

Advisors: Michael Rosenblum, Ravi Varadhan, Vadim Zipunnikov

Committee: Albert Wu, Sumithra Mandrekar

Alternates: Elizabeth Ogburn, Casey Rebholz

Acknowledgments

First of all, I am filled with gratitude for my primary advisor, Dr. Michael Rosenblum, who has advised me ever since I was admitted into this PhD in Biostatistics program. Michael, you mean a lot more than an advisor to me. In my first year, I was so afraid to fail the comprehensive exam since I did not have a solid math background. You cheered me up, you gave me confidence and you were the first to congratulate me on the best performance award in the comprehensive exam I received. You made me grow. You care a lot about your students. When I told you about my health issue, you kept that in mind, did many searches and forwarded me many helpful information. When you knew about my upcoming surgery, you asked me to put all the work off and relax, you also wondered if I have difficulty paying medical bills. You even marked my surgery on your busy calendar, sending your best wishes and giving encouragements around those days. You taught me techniques to care for baby just in time when I got struggled with my newborn. Academically, you brought me into the field of clinical trials, and I feel very lucky working with you on very interesting problems in this field. Whenever I got stuck

in the research, you were always there to lead me through. I learned a lot from you on the subject expertise, writing, presentation skills, and so on.

I would like to thank my two co-advisors, Dr. Ravi Varadhan and Dr. Vadim Zipunnikov. Ravi, I really enjoy the meetings with you every week, where, apart from academic discussion, we talk a lot on philosophical topics and life. You make me very productive. I knew you in the middle of my third year and just within a year, we had four papers planned under way (one is published, one is currently under review, one has a ready draft and another one is about to start). When one of the papers got rejected, I could not forget how confident you were in arguing against the referee in an appealing and it turned out the referee arguments were untenable. Vadim, thank you for introducing me into the world of wearable computing and an interesting project where we aim to monitor patient's recovery using actigraphy data. Thank you for being very kind and patient to me, to someone who barely knows what wearable computing does.

I wanted to thank Dr. Sumithra Mandrekar and Dr. Jun Yin for advising me during the internship at Mayo Clinic in summer 2016, where we produced two papers (under review) on Phase I study design and a published R package. Thank you for providing me with financial sponsorship for my fifth year in the PhD program, and fortunately we can continue to collaborate on the extension of our previous work.

Also, I would like to thank Dr. Alan Chiang and Dr. Yong Lin for being my

supervisors while I was doing internship at Eli Lilly in summer 2017. Thanks for the interesting project you have prepared for me on adaptive seamless trial designs, where we have an R Shiny Application developed that can be used in real trial planning as well as a paper in progress. Thank you for the full time job offer I received upon finishing the internship and I knew you gave the management level lots of positive comments on my performance.

Thank Dr. Albert Wu, Dr. Michael Rosenblum, Dr. Ravi Varadhan, Dr. Vadim Zipunnikov, Dr. Sumithra Mandrekar, Dr. Elizabeth Ogburn, Dr. Casey Rebholz and Dr. Jodi Segal for being willing to serve on my thesis committee, most of whom were also on my preliminary exam committee. Thank you for your helpful comments and suggestions that would surely improve my research. I would also like to thank Dr. Gary Rosner for joining my Phase II studies and providing many useful comments to make our work in shape.

I have been inspired a lot from the faculty members in the department, including Dr. Mei-Cheng Wang, Dr. Karen Bandeen-Roche, Dr. Marie Diener-West, Dr. Hongkai Ji, Dr. Daniel Scharfstein, Dr. Constantine Frangakis, among others. Thank you for showing great examples in research, teaching and service.

Special thanks to my peers, including Tianchen Qian, Yuxin Zhu, Haoyu Zhang, Hong Zhang, among others. Tianchen, I was impressed that you were never annoyed by the bombardment of my questions on research, and always ready and patient to discuss with me. Yuxin, I remember that I always troubled you with

probability theory problems in the first year, and you spent quite a lot of time walking me through those problems. I have always been grateful for your help and admire your fast and rigorous math mind. Haoyu, thank you for organizing so many Ping-Pang nights and basketball nights where students and peers hang out together, a great opportunity for our friendship bonding and thank you for promoting student benefits as a departmental representative. Hong, as always, thank you for your helpful answer to my random research questions.

Thank Mary Joy Argo for coordinating every thing in the department, in particular, you reminded me many times not to forget about course registration. Thank you for your patience and also for providing grammar check on my papers. I still remember Brian once said "for anything you don't know, ask Mary Joy!"

Particularly, I am so grateful to my family, my wife Shuyuan Wu, my son Will Du, my parents Chunlin Du and Liying Xie, and my parents-in-law Zeyong Wu and Mingxiu Bu. Shuyuan, you are always by my side and very supportive of my career choice. I really appreciate the courage you had when you decided to marry someone who had no idea about his future five years ago, even before I was offered an admission into this PhD program. I could not imagine what I will be like without you. Thank you, my dear, Shuyuan. Hey Will, father thanks you for making me a father, and bettering my understanding about the responsibility I should take. I do enjoy the time watching you grow. Thank my parents for insisting on me studying abroad and for your constant supports. It has been seven

years, and I assure you that you made the right choice. Thank my parents-in-law for coming to here, taking care of my wife for her postpartum recovery and the newborn so that I am able to spare some time working on this dissertation.

In the very end, thank the department graduate program committee for offering me admission into this PhD in Biostatistics program at Johns Hopkins Bloomberg School of Public Health, which makes a valuable and unforgettable five-year experience in my life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Before introducing a new experimental drug / treatment / device into the market, clinical trials are required by the regulatory agencies, like the Food and Drug Administration of the United States, the European Medicines Agency, etc., to demonstrate the efficacy and the safety of the product on the patients. Clinical trials typically consist of four phases, from Phase I studies to Phase IV studies, with each phase being a separate trial. At the completion of a phase, a decision will be made by the regulatory agency regarding whether or not to allow the product to advance to the next phase. Phase I studies usually recruit small number of patients (20 - 100), which mainly focus on the assessment of the safety of the product. Phase II studies involve a larger number of patients, up to hundreds, and start exploring the efficacy of the product. Phase III studies give a thorough testing of the product efficacy in up to thousands of patients, strictly controlling

the rate of false positive findings. Most of the Phase II and III studies are blinded, randomized trials, where patients are assigned randomly to the experimental arm and to the placebo/control arm, and neither patients nor researchers know about these assignments. Once the product passes Phase III trials and is approved by the regulatory agency, it can make a debut in the market for consumer use. Phase IV studies are Post Marketing Surveillance Trials, where the real-world effectiveness and/or the safety of the drug is evaluated.

Novel statistical methods do play an imperative role in the designs and analytics of clinical trials, making trials more efficient and cost-effective. For example, leveraging the baseline information into the estimation of treatment effect can lead to desirable gain in precision, thus substantial reduction in sample size of patients required to complete a trial, a huge saving. Also, individuals may vary in their responses to treatment, and the treatment may only benefit a subset of the overall study population. Therefore, statistical methods are needed to take into account the patient heterogeneity and to facilitate personalized medicine. Ignoring patient heterogeneity easily causes dilution of the treatment effect, and creates unethical situations where patients who do not benefit from the treatment are treated. Additionally, the interpretation of overall effect for treatment may not be meaningful, and in fact can be misleading. In this dissertation, I dedicate Chapters 2 to 5 to the study of the statistical methods and applications developed for Phase I to III studies, which are key to determining whether a product can be marketed or not.

Phase I designs traditionally use the dose-limiting toxicity (DLT), a binary endpoint from the first treatment cycle, to identify the maximum-tolerated dose (MTD) assuming a monotonically increasing relationship between dose and efficacy. In Chapter 2, I establish a general framework for a multi-stage adaptive design where I jointly model a continuous efficacy outcome and continuous/quasi-continuous toxicity endpoints from multiple treatment cycles. The normalized Total Toxicity Profile (nTTP) is used as an illustration for quasi-continuous toxicity endpoints, and I replace DLT with nTTP to take into account multiple grades and types of toxicities. In addition, the proposed design accommodates non-monotone dose-efficacy relationships, and longitudinal toxicity data in effort to capture the adverse events from multiple cycles. Extensive simulations showed that the design had a high probability of making the correct dose selection and good overdose control across various dose-efficacy and dose-toxicity scenarios. Furthermore, the proposed design allows for early termination when all doses are too toxic. To our best knowledge, the proposed Phase I dual-endpoint dose-finding design is the first such study to incorporate multiple cycles of toxicities and a continuous efficacy outcome.

In adaptive enrichment designs based on accrued data, early stopping of a subpopulation with sufficient evidence of treatment efficacy, futility, utility or harm is allowed according to preplanned rules for modifying enrollment criteria, while the remaining subpopulations continue to be enrolled. In Chapter 3, I propose and

evaluate a two-stage, Phase II adaptive enrichment clinical trial design. Its goal is to determine whether future phase III (confirmatory) trials should be conducted, and if so, which population should be enrolled. The population selected for phase III enrollment is defined in terms of a disease severity score measured at baseline. I optimize the phase II trial design and analysis in a decision theory framework. The resulting design is compared to simpler designs in simulation studies. I also apply the designs to resampled data from a completed, phase II trial evaluating a new surgical intervention for stroke.

Most existing methods for constructing trial designs are limited to situations where patient outcomes are observed soon after enrollment. This is a major barrier to the use of such designs in practice, since for many diseases the outcome of most clinical importance does not occur shortly after enrollment. In Chapter 4 for Phase III studies, I provide a general framework for the adaptive enrichment designs with delayed outcome, where I use semiparametric, locally efficient estimators at each interim analysis to leverage information in baseline variables and short-term outcomes to improve precision. I then evaluate power, expected sample size, bias, variance, and mean squared error for our design and compare with a non-adaptive design and unadjusted estimator. I demonstrate the advantage of our proposed methods, through simulations of a real trial. I strongly control the familywise Type I error rate, asymptotically.

Chapter 5 presents a useful and general method for exploring treatment effect

heterogeneity in Phase III studies, through identification of treatment-covariate interactions honoring different hierarchy conditions. I construct a single-step l_1 norm penalty procedure that maintains the hierarchical structure of interactions in a sense that the treatment-covariate interaction term is included in the model only when either the covariate or the covariate and the treatment both have non-zero main effects. I explore several parameterization schemes with different constraints added to Lasso that enforce the hierarchical interaction restriction. I solve the resulting constrained optimization problem using a spectral projected gradient method. I compare our methods to the unstructured Lasso using simulation studies covering a variety of scenarios for treatment-covariate interactions. The simulations show that our methods yield more parsimonious models and outperform unstructured Lasso in terms of prediction performance, and in terms of the ability to correctly identify non-zero treatment covariate interactions. The superior performance of our methods are also corroborated by an application to a large randomized clinical trial data investigating a drug for treating congestive heart failure (N=2,569). Our methods can be applied to continuous, binary and time to event outcome, providing a well-suited approach with sufficient flexibility in terms of parameterization for doing secondary analysis in Phase III trials to analyze heterogeneity in treatment effect.

Chapter 6 provides a summary of this dissertation and discusses the areas for future research work.

Chapter 2 is adapted from the working paper "**Yu Du**, Jun Yin, Daniel J. Sargent, Sumithra J. Mandrekar. *An Adaptive Multi-Stage Phase I Dose-finding Design Incorporating Continuous Efficacy and Toxicity Data from Multiple Treatment Cycles.*" Chapter 3 is adapted from the working paper "**Yu Du**, Gary L. Rosner, Michael Rosenblum. *Phase II Adaptive Enrichment Design to Determine the Population to Enroll in Phase III Trials, by Selecting Thresholds for Baseline Disease Severity.*" Chapter 4 is adapted from the working paper "**Yu Du**, Tianchen Qian, Huitong Qiu, Michael Rosenblum. *Bias, Variance, and Sample Size Reductions due to Adjustment for Prognostic Baseline Variables and Short Term Outcomes in Adaptive Enrichment Trial Designs with Delayed Outcomes.*" Chapter 5 is adapted from the working paper "**Yu Du**, Ravi Varadhan. *Lasso Estimation of Hierarchical Interactions for Analyzing Heterogeneous Treatment Effect.*"

Chapter 2

Phase I Studies: An Adaptive, Multi-Stage Dose-finding Design

SUMMARY.[1] Phase I designs traditionally use the dose-limiting toxicity (DLT), a binary endpoint from the first treatment cycle, to identify the maximum-tolerated dose (MTD) assuming a monotonically increasing relationship between dose and efficacy. In this chapter, we establish a general framework for a multi-stage adaptive design where we jointly model a continuous efficacy outcome and continuous / quasi-continuous toxicity endpoints from multiple treatment cycles. The normalized Total Toxicity Profile (nTTP) is used as an illustration for quasi-continuous toxicity endpoints, and we replace DLT with nTTP to take into account multiple grades and types of toxicities. In addition, the proposed design accommodates

[1] This Chapter 2 is adapted from the working paper "**Yu Du**, Jun Yin, Daniel J. Sargent, Sumithra J. Mandrekar. *An Adaptive Multi-Stage Phase I Dose-finding Design Incorporating Continuous Efficacy and Toxicity Data from Multiple Treatment Cycles.*"

non-monotone dose-efficacy relationships, and longitudinal toxicity data in effort to capture the adverse events from multiple cycles. Stage 1 of our design uses toxicity data to perform dose-escalation and identify a set of initially allowable (safe) doses; stage 2 of our design incorporates an efficacy outcome to update the set of allowable doses for each new cohort and randomizes the new cohort of patients to the allowable doses with emphasis towards those with higher predicted efficacy. Stage 3 uses all data from all treated patients at the end of the trial to make final recommendations. Simulations showed that the design had a high probability of making the correct dose selection and good overdose control across various dose-efficacy and dose-toxicity scenarios. In addition, the proposed design allows for early termination when all doses are too toxic. To our best knowledge, the proposed dual-endpoint dose-finding design is the first such study to incorporate multiple cycles of toxicities and a continuous efficacy outcome.

2.1 Background

Phase I clinical trials are designed to identify the recommended Phase II dose (RP2D) for future trials. For cytotoxic agents, the recommended dose is generally the maximum-tolerated dose (MTD). Traditionally, grade 3 or 4 toxicity events in the first treatment cycle are defined as dose-limiting toxicity (DLT) events and used to determine MTD. Late cycle toxicity events are recorded but not used in the

dose assessment. The current development of molecularly targeted agents (MTAs) poses new challenges for the study design of phase I cancer clinical trials. Unlike the cytotoxic agents that are often administrated for a limited number of treatment cycles, MTAs are administered until disease progression and often have a very different toxicity profile, characterized by chronic, prolonged events or cumulative toxicity as opposed to the early onset of severe adverse events. Therefore, it becomes important to consider toxicity events after the first treatment cycle, as well as repeated and chronic occurrence of lower grade events besides just grade 3 and 4 events. The idea of using longitudinal toxicity information have been explored by Legedza and Ibrahim (2000), Braun et al. (2007) and Yin et al. (2017).

MTAs also have different toxicity-efficacy relationships compared to cytotoxic agents. Numerous studies have identified that lower doses of MTAs may offer similar efficacy as higher doses. As a result, searching for the MTD may not be the optimal treatment strategy for MTAs. Incorporation of early efficacy signals if available has become important for these novel agents. Only limited methods exist for dose-finding designs that account for toxicities and continuous efficacy as dual-endpoints in the selection of recommended Phase II dose. Thall and Cook (2004) used a set of efficacy-toxicity trade-off contours to accommodate either trinary or bivariate binary outcomes that include efficacy. Bekele and Shen (2005) incorporated the correlation between the binary toxicity and continuous activity outcome via a latent Gaussian variable. Houede et al. (2010) used a generalization

of the Aranda-Ordaz model and a Gaussian copula to model the marginal outcome and joint distribution of dual toxicity and efficacy endpoints. A common feature of these designs is that they only considered binary toxicity endpoints in Phase I dose-finding designs, ignoring the fact that toxicity data are high dimensional in nature, with various types, grades, attribution over multiple treatment cycles. Bekele and Thall (2004) proposed the total toxicity burden (TTB) as the arithmetic sum of different grades and types of toxicities, weighted by the severity weights elicited from clinicians. Lee et al. (2012) proposed the toxicity burden score (TBS) to summarize toxicity using a weighted sum, where the severity weights were estimated via regression using historical data. Ezzalfani et al. (2013) proposed another flexible toxicity endpoint, called the total toxicity profile (TTP), upon which our proposed method is based as an illustration. However, the proposed design can easily be generalized to any continuous/quasi-continuous toxicity endpoints, including TTB, TBS, and etc.

The total toxicity profile (TTP) was developed to overcome the oversimplification of toxicity data in phase I trials (Ezzalfani et al., 2013). The TTP score captures multiple types and grades of toxicities occurring during the first treatment cycle. Following the notation in Ezzalfani et al. (2013), let w_{lh} denote the elicited weight of toxicity type l ($l \in \{1, \ldots, L\}$) occurring at grade h ($h \in \{0, \ldots, 4\}$). Hence, the weight vector for toxicity l is $\boldsymbol{w_l} = (w_{l0}, \ldots, w_{l4})'$ and the weight matrix is denoted as $W = (\boldsymbol{w_1}, \ldots, \boldsymbol{w_L})'$. For patient i, denote the maximum observed grade

of toxicity type l as G_{il}. Then the TTP_i is defined as

$$TTP_i = \sqrt{\sum_{l=1}^{L} \sum_{h=0}^{4} w_{lh}^2 1(G_{il} = h)}, \tag{2.1}$$

where $1(G_{il} = h)$ is an indicator function which takes value 1 if the maximum observed grade is h for toxicity type l, and 0 otherwise. The TTP is further normalized to nTTP, in order to constrain the continuous toxicity endpoint to be within 0 and 1, $nTTP_i = \frac{TTP_i}{v}$, where v is a normalization constant (for details, please refer to Ezzalfani et al. (2013)). This is an improvement from using the traditional DLT based endpoint in that it takes into account clinical multidimensionality of multiple types/grades of toxicities for a given toxicity profile, as well as the moderate toxicity events commonly ignored by using DLT endpoint. However, patients participating in Phase I clinical trials usually receive more than one cycle of experimental regimen in the absence of DLT or disease progression. To account for this limitation, an extension to this design to incorporate nTTP scores from multiple treatment cycles was proposed by Yin et al. (2017).

We propose a three-stage dose-finding design based on both longitudinal toxicity data of nTTP scores and an efficacy outcome. This is an extension of our previous work (Yin et al., 2017), which only uses nTTP scores from multiple treatment cycles. Furthermore, the proposed design is easily extended to other continuous / quasi-continuous toxicity endpoints than nTTP. To our best knowledge, it is the

first design to simultaneously account for both the multiple types/grades toxicity events over multiple treatment cycles and a continuous efficacy outcome. In Section 2.2, a joint model is introduced that takes into account the correlation between the toxicity data over multiple treatment cycles and a continuous efficacy outcome. Section 2.3 describes a three-stage dose-finding algorithm where stage 1 of the design uses toxicity data to perform dose-escalation and to identify a set of initially allowable (safe) doses; stage 2 of the design incorporates an efficacy outcome to update the set of allowable doses and randomizes the new cohort of patients to the allowable doses with emphasis towards those with higher predicted efficacy; stage 3 uses all data from all treated patients at the end of the trial to make final recommendations. The recommended Phase II dose is, as we suggest, the lowest allowable dose with acceptable efficacy, although the final decision of dose selection is left to the discretion of the physicians upon completion of the trial, based on the estimated profile of efficacy and toxicities the method outputs. In Section 2.4, extensive simulations are conducted to evaluate the operating characteristics of the proposed design under various clinical relevant scenarios. Section 2.5 concludes the chapter with a discussion of the proposed design, as well as some insights into future work.

2.2 Joint Model

We jointly model the toxicity profile, nTTP scores across multiple cycles, and a continuous efficacy outcome for each patient. The joint model, introduced by Wang et al. (2000), is comprised of two submodels: for longitudinal nTTP scores, a linear mixed effect model; while for continuous efficacy outcome, a linear model. Here, we transform any continuous efficacy outcome so that it ranges from 0 to 1 by using empirical distribution function. Suppose we have cumulative data for n patients. Let Y_{ij} be the nTTP score for patient i ($i = 1, 2, \ldots, n$) at cycle j, where $t_{ij} = t_j = j$ represents the j^{th} ($j = 1, 2, \ldots, J$) cycle of treatment, the same across all patients. We use x ($x = 1, 2, \ldots, K$) to denote the dose of the agent and x_i is the dose allocated to patient i. We assume that the same dose is given to a patient in each cycle of the treatment. Let E_i be the continuous efficacy outcome for the i^{th} patient.

The linear mixed effect submodel with a random intercept for the longitudinal nTTP scores is such that

$$Y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 t_j + u_i + \epsilon_{ij}, \tag{2.2}$$

where

$$\epsilon_{ij} \sim N(0, \sigma_\epsilon^2),$$

$$u_i \sim N(0, \sigma_u^2),$$

and the linear submodel for the efficacy outcome is given below:

$$E_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \gamma u_i + \epsilon_{e_i}, \tag{2.3}$$

where

$$\epsilon_{e_i} \sim N(0, \sigma_e^2).$$

We assume that $u_1, u_2, \ldots, u_n$, $\epsilon_{11}, \epsilon_{12}, \ldots, \epsilon_{nJ}$, $\epsilon_{e_1}, \epsilon_{e_2}, \ldots, \epsilon_{e_n}$ are independent where u_i is the random intercept for the i^{th} patient, shared with the two submodels, ϵ_{ij} is the measurement error for patient i nTTP score at cycle j and ϵ_{e_i} is the measurement error for efficacy outcome for patient i. Let θ represent the parameters to be estimated such that $\theta = (\beta_0, \beta_1, \beta_2, \alpha_0, \alpha_1, \alpha_2, \gamma, \sigma_\epsilon, \sigma_u, \sigma_e)'$.

Let us use $\pi_Y(x, \theta, t_j)$ to denote the mean of nTTP score for a randomly selected patient who undergoes the j^{th} treatment cycle with allocated dose level x given the parameters θ, so that $\pi_Y(x, \theta, t_j) = \beta_0 + \beta_1 x + \beta_2 t_j$. Therefore, $\beta_0, \beta_1, \beta_2$ relate to the mean structure of toxicity profile. In order to reflect the fact that the toxicities of most cancer agents increase with dose, we add a constraint a priori to the dose effect β_1 such that $\beta_1 > 0$. Also, let $\pi_E(x, \theta)$ represent the mean of efficacy outcome for the patient to whom dose level x is allocated given the parameters θ, so that $\pi_E(x, \theta) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$. Hence, $\alpha_0, \alpha_1, \alpha_2$ maneuver a quadratic

mean structure of efficacy outcome. This quadratic mean structure, $\pi_E(x, \theta)$, is capable of describing most common patterns between efficacy and dose, including a flat (the efficacy does not change with dose), increasing/decreasing (the efficacy increases/decreases as dose level goes up), plateau (the efficacy increases and reaches a plateau as dose level increases) and a parabolic (the efficacy first increases and then decreases as dose level varies from low to high) relationship. We will explore these scenarios in Section 2.4. $\sigma_\epsilon, \sigma_e$ are standard deviations of the measurement errors for nTTP score and efficacy outcome, respectively while σ_u is the standard deviation of random intercepts. Random intercepts u_i serve as the linkage between two submodels, and γ indicates the strength of the association such that the covariance between the nTTP score and efficacy outcome for a patient, given the treatment cycle, the parameters θ and the allocated dose level, is $\gamma\sigma_u^2$.

A patient will discontinue the treatment anytime they experience a DLT. Hence, patients can have difference number of treatment cycles. However, efficacy outcome is assumed to be always measured at a fixed time point during the treatment (e.g., at the end of cycle 3), patients who drop out due to DLT before that time point will have missing efficacy data.

In the early stages of our design, we do not have patients' toxicity/efficacy information for all J cycles of treatment. We, therefore, introduce a simplified

version of submodel (2.2), as below:

$$Y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 t'_j + u_i + \epsilon_{ij}, \tag{2.4}$$

where t'_j dichotomizes t_j, a binary indicator of first cycle or beyond, such that

$$t'_j = \begin{cases} 0 & \text{if } j = 1 \\ 1 & \text{if } j = 2, \ldots, J, \end{cases}$$

where all other components are the same. This, in addition, emphasizes our concern about the toxicity at the first cycle of treatment and the toxicities for the other cycles beyond as an aggregate measure.

2.2.1 Estimation

For parameter estimation of the joint model, we implement Bayesian methods, Markov Chain Monte Carlo (MCMC) to estimate the posterior distribution of the parameters, using JAGS (Plummer, 2003). Non-informative prior distributions are specified since elicitation in this context is difficult, therefore, the estimation of the parameters will largely depend on data. For parameters $\beta_0, \beta_2, \alpha_0, \alpha_1, \alpha_2, \gamma$, we specify independent normal priors with mean 0 and precision 0.001 (the inverse of the variance). For all other parameters, $\sigma_\epsilon, \sigma_u, \sigma_e$, we use uniform priors from 0

to 1000. We discard the first $5,000$ samples as burn-in and sample the subsequent $5,000$ iterations. The posterior mean of $\pi_Y(x, \theta, t_j)$ and $\pi_E(x, \theta)$ are used to guide the dose-escalation/de-escalation which will be specified in Section 2.3.

2.3 Dose-finding Algorithm

A three-stage Phase I dose-finding algorithm, whose goal is to find efficacious doses that are also safe, is proposed. Suppose N is the maximum sample size. Stage 1 only utilizes longitudinal data of nTTP scores to navigate the dose-escalation for each subsequent patient cohort of size m as well as identifies initially allowable (safe) doses in terms of toxicity when cumulative sample size reaches $\frac{N}{2}$. Stage 2 involves efficacy outcome, and by jointly modeling the toxicity and efficacy as specified in Section 2.2, it keeps updating the set of allowable doses and randomizes patients to the allowable doses with emphasis towards those with higher predicted efficacy for each new cohort. Stage 3 concludes the algorithm when all the data from all treated patients becomes available (i.e., at the end of the trial), where we fit the joint model to the full data and find the efficacious doses that are also allowable (safe). Early termination of the trial is allowed, when there are no allowable (safe) doses in terms of toxicity but skipping of dose levels not previously tried is not allowed at any time. Below we specify how each stage works in more details.

2.3.1 Stage 1

Stage 1 aims to perform the dose-escalation and define initially allowable (safe) doses in terms of toxicity, based on longitudinal data of nTTP scores only until the sample size of patients enrolled reaches $\frac{N}{2}$. We enroll patients by cohorts of size m.

Dose 1 is assumed to be the starting dose and from the first cohort of patients, we use the $3+3$ design for dose escalation until we escalate to dose 2. Once dose 2 is allocated to a new cohort and when its first cycle nTTP scores become available, we switch to the linear mixed effect submodel (2.4) in Section 2.2, fitting the cumulative longitudinal toxicity data for dose escalation. We employ the submodel (2.4) when we have toxicity data for more then one dose level, in order to give more accurate estimates than applying the model with only one dose level.

Suppose $\mathbb{O}$ denotes the current cumulative data and based on $\mathbb{O}$, the dose-escalation is guided by Bayesian risk evaluation. We define a loss function L for dose level x and cycle 1 of the treatment such that

$$L\left(\pi_Y(x,\theta,t_1'),\pi_1\right) = \left|\pi_Y(x,\theta,t_1') - \pi_1\right|,$$

where π_1 is the target toxicity in nTTP score in the first cycle of treatment, and $\pi_Y(x,\theta,t_1')$ is the mean nTTP score given parameter θ, at cycle 1 of the treatment for dose level x. This loss function L was used by Yin et al. (2017). The dose level x that delivers the minimum $\mathbb{E}\left[L\left(\pi_Y(x,\theta,t_1'),\pi_1\right)|\mathbb{O}\right]$, namely minimizing the

Bayesian risk, will be allocated to the next cohort of patients. Once the first cycle toxicity data becomes available for the third enrolled cohort of patients, the dose x will be declared an allowable (safe) dose based on safety if the following two conditions are met:

$$\mathbb{P}\{\pi_Y(x, \theta, t_1') < c_1 | \mathbb{O}\} > p_1, \tag{2.5}$$

$$\mathbb{P}\{\pi_Y(x, \theta, t_j' = 1) < c_2 | \mathbb{O}\} > p_2, \tag{2.6}$$

where c_1, c_2 are the upper bounds of the mean nTTP score for the first and subsequent cycles of treatment, while p_1, p_2 are the corresponding probability cutoffs respectively. These parameters need to be pre-defined, and should be chosen based on both physicians' discretion and design's operating characteristics, which will be discussed in Section 2.5. Condition (2.5) concerns the toxicity at the first cycle of treatment, while condition (2.6) regards the toxicities for the other cycles, with $\pi_Y(x, \theta, t_j' = 1)$ serving as an aggregate mean nTTP score for treatment cycles beyond cycle 1 at dose level x, given the parameter θ. These two probability conditions (2.5) and (2.6) act here as a stopping rule such that the trial shall be terminated early if there is no allowable (safe) dose. Otherwise, stage 1 continues until the cumulative sample size of patients enrolled reaches $\frac{N}{2}$ and at the end of stage 1, we identify a set of initially allowable doses, $\mathbb{A}$, ready to enter stage 2.

2.3.2 Stage 2

Once the efficacy data becomes available on patients from stage 1, we fit the joint model (2.3) and (2.4), and update the set of allowable (safe) doses, $\mathbb{A}$, using the two probability conditions (2.5), (2.6). The next cohort of patients are randomized to the allowable doses with emphasis towards those dose levels with higher predicted efficacy. The probability that dose level $a \in \mathbb{A}$ is assigned is as follows:

$$\frac{\exp\left\{\mathbb{E}\left[\pi_E(a,\theta)|\mathbb{O}\right]\right\}}{\sum_{x\in\mathbb{A}}\exp\{\mathbb{E}\left[\pi_E(x,\theta)|\mathbb{O}\right]\}},$$

where $\mathbb{E}\left[\pi_E(a,\theta)|\mathbb{O}\right]$ is the posterior mean of efficacy outcome, the predicted efficacy, for dose level $a \in \mathbb{A}$. Thus, the higher the predicted efficacy of a dose, the more likely it will be assigned to the next cohort of patients. We continue this update for the set of allowable doses, $\mathbb{A}$, and efficacy-based randomization for each new cohort until the maximum sample size N is reached. Again, early termination is allowed if there is no allowable (safe) dose at any time in this stage, i.e., $\mathbb{A} = \emptyset$.

2.3.3 Stage 3

The purpose of Stage 3 is to define the efficacious doses that are also safe at the end of the trial and make a final recommendation of the dose level for future, Phase II trial. When all data from all treated patients become available, we implement the joint modeling (2.2) and (2.3) as described in Section 2.2, a fully parameterized

version.Again, we update the set of allowable doses $\mathbb{A}$ using the two probability conditions (2.5) and (2.6), with t'_j replaced by t_j, and for the second condition (2.6) regarding later cycles of treatment beyond cycle 1, we use instead

$$\mathbb{P}\{\overline{\pi_Y(x, \theta, t_j > 1)} < c_2 | \mathbb{O}\} > p_2, \tag{2.7}$$

where $\overline{\pi_Y(x, \theta, t_j > 1)} = \frac{1}{J-1} \sum_{j=2,\ldots,J} \pi_Y(x, \theta, t_j = j)$. We aggregate the toxicity information for cycles 2 through J by taking the average. Subsequently, for each allowable dose $a \in \mathbb{A}$, we compute the posterior mean of efficacy response, $\mathbb{E}[\pi_E(a, \theta)|\mathbb{O}]$. Let $l \in \mathbb{A}$ be the allowable dose that gives the largest posterior mean of efficacy, such that $l = \text{argmax}_{a \in \mathbb{A}} \mathbb{E}[\pi_E(a, \theta)|\mathbb{O}]$. We define a proximity threshold δ such that any dose $a \in \mathbb{A}$ that satisfies the condition $|\mathbb{E}[\pi_E(a, \theta)|\mathbb{O}] - \mathbb{E}[\pi_E(l, \theta)|\mathbb{O}]| \leq \delta$ will be declared an efficacious dose since it is in the efficacy proximity of the dose l, that gives the largest posterior mean of efficacy. We thus define the set of efficacious doses that are also safe, $\mathbb{H}$, such that $\mathbb{H} = \{h \in \mathbb{A} : |\mathbb{E}[\pi_E(h, \theta)|\mathbb{O}] - \mathbb{E}[\pi_E(l, \theta)|\mathbb{O}]| \leq \delta\}$.

We report the set $\mathbb{H}$ at the end of the trial along with the toxicity profile from different cycles of the treatment, $\mathbb{E}[\pi_Y(h, \theta, t_j)|\mathbb{O}]$, and efficacy profile, $\mathbb{E}[\pi_E(h, \theta)|\mathbb{O}]$, for any $h \in \mathbb{H}$. Any dose $h \in \mathbb{H}$ can be picked based on physician's judgment and preference but in our work, we suggest the lowest (safest) dose in set $\mathbb{H}$ as the final recommendation for future, Phase II trial. We explore several scenarios with differ-

ent toxicity and efficacy structures, and conduct simulation studies in Section 2.4 to evaluate the performance of operating characteristics for the design.

2.4 Simulation Studies

2.4.1 Generation of the Toxicity Data

Let us assume that there are three independent types of toxicities related to the treatment: renal, neurological and hematological toxicities. The clinical weight matrix, elicited from the physician, is shown as follows:

$$W = \begin{pmatrix} 0 & 0.5 & 0.75 & 1 & 1.5 \\ 0 & 0.5 & 0.75 & 1 & 1.5 \\ 0 & 0 & 0 & 0.5 & 1 \end{pmatrix}.$$

DLT, in this setting, is defined as the occurrence of a grade 3 or 4 renal / neurological toxicity or a grade 4 hematological toxicity. Based on this weight matrix, we compute the TTP score for patients. The maximum TTP score that can be acquired from the matrix is 2.34, corresponding to the scenario where a patient experiences grade 4 toxicities for all three types. We thus normalize TTP by dividing the value by a normalization constant, 2.5, to obtain nTTP for each patient. The target toxicity nTTP elicited from the physician is set at 0.28. For details on the normalization constant and the target nTTP at cycle 1 of the treatment, please refer to Ezzalfani

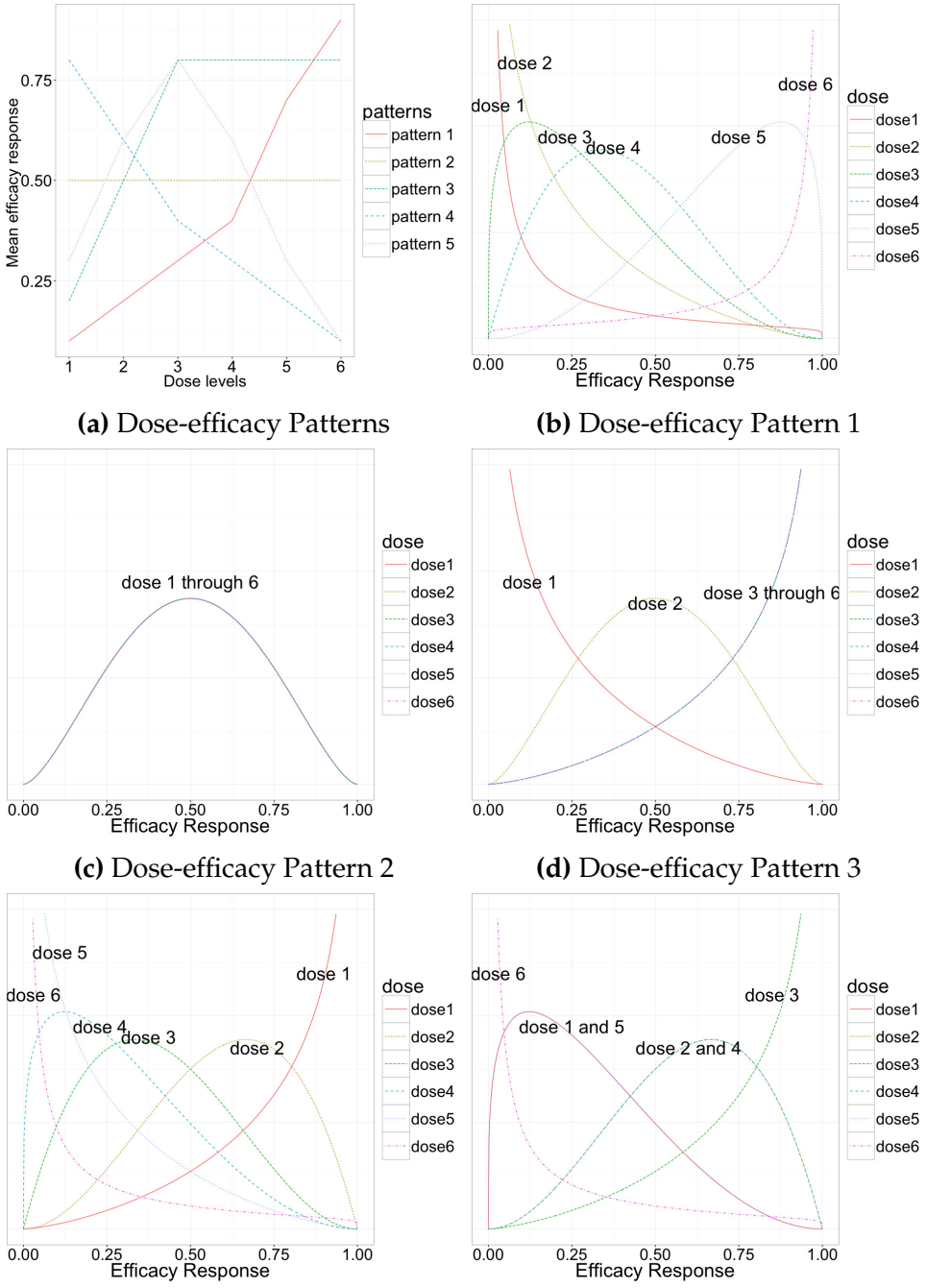
et al. (2013). In the simulation, however, we need to simulate the toxicity profile for each patient in order to compute an nTTP score. Please see Appendix for more details on how toxicity data is generated non-parametrically.

2.4.2 Generation of the Efficacy Outcome

The efficacy outcome, in our setting, is assumed to be a continuous variable from 0 to 1. We explore 5 common dose-efficacy patterns in the simulation study, as described in Section 2.2. We generate efficacy outcome for each patient from the beta distribution. Figure 2.1 displays the dose-efficacy patterns and the distribution of efficacy outcome for each dose in each pattern. It is clearly seen from dose-efficacy pattern 1 to 5 that efficacy data distributions have large variation and/or skewness. For more details, please refer to Appendix.

2.4.3 Simulation Settings

We consider 14 scenarios in total, that takes into account 6 toxicity structures and 5 dose-efficacy patterns. For each scenario, we simulated 1,000 trials that explore 6 dose levels with maximum 6 treatment cycles, where dose level 1 is the starting dose. We enroll patients by cohorts of 3 and the maximum sample size for each trial is set at 36. We assume that efficacy outcome is measured at the end of cycle 3 of the treatment so if the patient drops out of the trial before cycle 3, his/her



(a) Dose-efficacy Patterns

(b) Dose-efficacy Pattern 1

(c) Dose-efficacy Pattern 2

(d) Dose-efficacy Pattern 3

(e) Dose-efficacy Pattern 4

(f) Dose-efficacy Pattern 5

Figure 2.1: Dose-efficacy patterns and the distribution of efficacy outcome for each dose in each pattern.

efficacy data will be missing. Missing information is also present for toxicity data due to DLT dropouts.

- Parameters
  - $\pi_1 = 0.28$, the target toxicity nTTP for cycle 1 of the treatment is set at 0.28.
  - $c_1 = c_2 = \pi_1$, the upper bounds of toxicity for cycle 1 and the subsequent cycles are set equal to π_1.
  - $p_1 = p_2 = 0.2$, the two probability cutoffs for conditions (2.5) and (2.6) are set at 0.2.
  - $\delta = 0.1$, so that all the safe doses whose posterior means of efficacy are within 0.1 away from the largest one among the safe doses will be declared efficacious.
- Performance metrics
  - $\mathbb{P}_{\text{recom}}$, the percentage of dose recommendation, where we recommend the lowest dose that is efficacious.
  - $\mathbb{P}_{\text{alloc}}$, the percentage of dose allocation, computed by dividing the number of patients allocated to the dose by the total number of patients enrolled.
  - $\mathbb{P}_{\text{effy}}$, the percentage of dose being efficacious and safe.

Note that $\mathbb{P}_{\text{effy}}$ is often different from $\mathbb{P}_{\text{recom}}$ since $\mathbb{P}_{\text{effy}}$ indicates the probability that any particular dose is selected as an efficacious dose that is also safe while $\mathbb{P}_{\text{recom}}$ encodes the probability that a dose is recommended as the lowest dose that is also efficacious. There may be multiple safe and efficacious doses for each trial, among which we suggest the lowest dose in the study, though in practice, the final recommendation is at physician's discretion. The simulation study is performed using R (R Core Team, 2016).

2.4.4 Results

The simulation results for scenario 1 to 5 are summarized in Table 2.1. Scenario 1 to 5 correspond to the dose-efficacy pattern 1 to 5 in Figure 2.1, i.e., the increasing, flat, plateau, decreasing and parabolic relationship for efficacy with dose, respectively. In addition, they have a toxicity structure where the target toxicity dose (MTD) at cycle 1 of the treatment is dose 5 and the cycle effect is flat, namely, the mean nTTP (mnTTP) for each dose level is the same across cycles.

In scenario 1, the most efficacious dose among the safe doses is dose 5 whose mean of efficacy is 0.7. seventy-seven percent of the trials declare this dose as an efficacious dose and 72% of the recommendations correspond to this dose. Dose 4 is the second best choice in this scenario if physician prefers a lower toxicity, which shares 11.1% chance of the recommendations. In total, 83.1% of the recommendations go to these top 2 choices. Scenario 2 has a flat relationship in terms

Table 2.1: Simulation results for scenario 1 to 5: $\mathbb{P}_{\text{alloc}}$, the percentage of dose allocation, $\mathbb{P}_{\text{recom}}$, the percentage of dose recommendation, where we recommend the lowest (safest) dose that is also efficacious, and $\mathbb{P}_{\text{effy}}$, the percentage of dose being efficacious and safe. mnTTP represents the mean nTTP score while pDLT refers to the probability of DLT event. The target toxicity dose at cycle 1 of the treatment is dose 5, with flat cycle effect.

	Dose Levels						
	1	2	3	4	5	6	None
mnTTP(cycle 1)	0.046	0.054	0.108	0.183	0.280	0.359	
mnTTP(cycle 2 through 6)	0.046	0.054	0.108	0.183	0.280	0.359	
pDLT	0.008	0.011	0.064	0.195	0.330	0.446	
			Scenario 1				
Mean Efficacy	0.1	0.2	0.3	0.4	0.7	0.9	
$\mathbb{P}_{\text{alloc}}$	16.4	17.3	18.2	21.1	20.9	6.2	
$\mathbb{P}_{\text{recom}}$	1.3	0.1	1.8	11.1	72.0	13.7	0.0
$\mathbb{P}_{\text{effy}}$	1.3	0.2	2.1	13.7	77.7	15.5	0.0
			Scenario 2				
Mean Efficacy	0.5	0.5	0.5	0.5	0.5	0.5	
$\mathbb{P}_{\text{alloc}}$	18.2	18.3	19.0	20.7	18.2	5.5	
$\mathbb{P}_{\text{recom}}$	68.5	17.8	6.2	2.1	3.3	2.1	0.0
$\mathbb{P}_{\text{effy}}$	68.5	71.1	69.8	69.6	62.4	11.3	0.0
			Scenario 3				
Mean Efficacy	0.2	0.5	0.8	0.8	0.8	0.8	
$\mathbb{P}_{\text{alloc}}$	15.0	17.6	20.0	22.9	19.2	5.4	
$\mathbb{P}_{\text{recom}}$	0.0	0.2	41.1	53.6	4.5	0.6	0.0
$\mathbb{P}_{\text{effy}}$	0.0	0.2	41.3	94.5	79.5	8.3	0.0
			Scenario 4				
Mean Efficacy	0.8	0.6	0.4	0.3	0.2	0.1	
$\mathbb{P}_{\text{alloc}}$	21.7	19.7	18.0	19.0	16.7	4.9	
$\mathbb{P}_{\text{recom}}$	99.7	0.2	0.0	0.0	0.0	0.1	0.0
$\mathbb{P}_{\text{effy}}$	99.7	12.7	2.8	0.1	0.0	0.4	0.0
			Scenario 5				
Mean Efficacy	0.3	0.6	0.8	0.6	0.3	0.1	
$\mathbb{P}_{\text{alloc}}$	16.5	19.6	21.0	21.7	16.7	4.7	
$\mathbb{P}_{\text{recom}}$	3.2	40.8	55.9	0.0	0.0	0.0	0.1
$\mathbb{P}_{\text{effy}}$	3.2	43.9	99.7	58.5	0.8	0.0	0.1

of dose-efficacy pattern since all doses have a mean of efficacy equal to 0.5. About 69% of the recommendations correspond to the lowest dose, dose 1, which is the optimal one. Around 18% of simulated trials recommend dose 2, which is equally desirable as dose 1 since they have very close toxicity across all cycles, 0.046 (dose 1) vs 0.054 (dose 2). Nearly 87% of the trials, in total, make a recommendation for dose 1 or dose 2. Scenario 3 demonstrates a plateau relationship between efficacy and dose, where the efficacious doses, by its definition, are dose 3 to 5. When doses have comparable efficacy, the lower doses will be preferred since they give lower toxicities. Almost 95% of the recommendations correspond to dose 3 or dose 4, which are lower doses than dose 5 yet have the same efficacy. Almost 100% of the trials recommend dose 1, the most efficacious dose, in scenario 4, where the mean of efficacy decreases as dose level varies from low to high. Scenario 5 shows a parabolic relationship for dose-efficacy pattern and 99.7% of the trials declare that dose 3, whose mean of efficacy is at the peak, is an efficacious dose. Nearly 97% of the recommendations come to the top 2 choices, dose 3 and dose 2.

Table 2.2 displays the simulation results for scenario 6 to 10, where the only change from scenario 1 to 5 is that target toxicity dose is dose 3 instead of dose 5, so that dose 4 to 6 are not considered safe doses. In scenario 6, dose 2 and 3, by definition, are efficacious doses since their mean of efficacy are within $\delta = 0.1$ distance. Around 70% of the recommendations correspond to the efficacious doses. Again, in a flat relationship between efficacy and dose, in scenario 7, 95.4% of the

trials recommend top 2 choices, dose 1 or dose 2, which are lower doses than dose 3. Scenario 8 behaves similarly as scenario 10, since when we focus on the safe doses, dose 1 to 3, they have a similar increasing dose-efficacy pattern in both of these two scenarios. Around 96% of the recommendations go for the two best doses, dose 3 or dose 2 in both scenario 8 and 10. Scenario 9 resembles scenario 4, where 99.2% of the trials recommend the best dose, dose 1, that has the highest mean of efficacy.

In terms of the metric $\mathbb{P}_{\text{alloc}}$, all scenarios 1 to 10 allocate less than 6.2% of the patients to the doses more toxic than the target toxicity dose, which demonstrates good overdose control of our design.

The simulation results for the last four scenarios, scenario 11 to 14, are demonstrated in Table 2.3, where we fix a dose-efficacy pattern, the parabolic relationship. Scenario 11 demonstrates a case where all doses are toxic with mnTTP larger than the target toxicity, 0.28. None of the doses are recommended in this scenario, with 88% of the patients being allocated to the starting dose, dose 1. Since our design allows early termination if all doses are too toxic, the average number of patients enrolled is 10. Therefore, in scenario 11, most of the trials terminate early after enrolling 3 cohorts of patients. Scenario 12 corresponds to the case where only dose 1 is safe and our design recommends dose 1 for 94.6% of the simulated trials. Scenario 13 and 14 show the increasing cycle effect and the decreasing cycle effect respectively, where the mnTTP for each dose increases/decreases as patients go

Table 2.2: Simulation results for scenario 6 to 10: $\mathbb{P}_{\text{alloc}}$, the percentage of dose allocation, $\mathbb{P}_{\text{recom}}$, the percentage of dose recommendation, where we recommend the lowest (safest) dose that is also efficacious, and $\mathbb{P}_{\text{effy}}$, the percentage of dose being efficacious and safe. mnTTP represents the mean nTTP score while pDLT refers to the probability of DLT event. The target toxicity dose at cycle 1 of the treatment is dose 3, with flat cycle effect.

	Dose Levels						
	1	2	3	4	5	6	None
mnTTP(cycle 1)	0.108	0.183	0.280	0.359	0.409	0.433	
mnTTP(cycle 2 through 6)	0.108	0.183	0.280	0.359	0.409	0.433	
pDLT	0.064	0.195	0.330	0.446	0.513	0.558	
			Scenario 6				
Mean Efficacy	0.1	0.2	0.3	0.4	0.7	0.9	
$\mathbb{P}_{\text{alloc}}$	28.4	40.5	26.6	4.3	0.2	0.0	
$\mathbb{P}_{\text{recom}}$	27.9	40.7	28.9	2.3	0.0	0.1	0.1
$\mathbb{P}_{\text{effy}}$	27.9	63.9	59.6	3.2	0.0	0.1	0.1
			Scenario 7				
Mean Efficacy	0.5	0.5	0.5	0.5	0.5	0.5	
$\mathbb{P}_{\text{alloc}}$	30.3	40.6	25.1	3.8	0.2	0.0	
$\mathbb{P}_{\text{recom}}$	75.1	19.3	3.9	1.3	0.1	0.3	0.0
$\mathbb{P}_{\text{effy}}$	75.1	75.2	43.0	2.3	0.2	0.3	0.0
			Scenario 8				
Mean Efficacy	0.2	0.5	0.8	0.8	0.8	0.8	
$\mathbb{P}_{\text{alloc}}$	25.9	41.6	28.1	4.2	0.2	0.0	
$\mathbb{P}_{\text{recom}}$	2.4	36.8	59.5	1.1	0.1	0.0	0.1
$\mathbb{P}_{\text{effy}}$	2.4	38.6	62.0	1.8	0.1	0.0	0.1
			Scenario 9				
Mean Efficacy	0.8	0.6	0.4	0.3	0.2	0.1	
$\mathbb{P}_{\text{alloc}}$	33.3	40.0	23.0	3.5	0.2	0.0	
$\mathbb{P}_{\text{recom}}$	99.2	0.7	0.0	0.0	0.0	0.0	0.1
$\mathbb{P}_{\text{effy}}$	99.2	15.6	0.8	0.3	0.0	0.0	0.1
			Scenario 10				
Mean Efficacy	0.3	0.6	0.8	0.6	0.3	0.1	
$\mathbb{P}_{\text{alloc}}$	25.8	43.1	26.7	4.0	0.3	0.1	
$\mathbb{P}_{\text{recom}}$	2.6	49.4	46.7	0.9	0.3	0.0	0.1
$\mathbb{P}_{\text{effy}}$	2.6	51.3	63.0	2.9	0.3	0.0	0.1

through later cycles. Both of these scenarios have similar results under parabolic dose-efficacy relationship: around 97% of the recommendations go for the best two doses, dose 3 or dose 2, while more than 99% of the trials declare that dose 3, at the efficacy peak, is an efficacious dose.

2.5 Remarks

Conventional dose-finding designs for Phase I clinical trials do not utilize efficacy data and toxicity data from late treatment cycles in the dose-finding decision. The rapid development of cancer immunotherapy and MTAs require the consideration of both efficacy and toxicity data that are often ignored in traditional dose-finding studies. As a result, new strategies must be developed for these Phase I dose-finding studies.

This chapter presents a novel Phase I design that combines an early-stage efficacy endpoint and toxicity profiles from multiple treatment cycles during the dose-finding process. It is an extension of our previous work (Yin et al., 2017) that only uses toxicity data from multiple treatment cycles. The proposed design uses a novel three-stage procedure in dose-finding. The advantage of the three-stage procedure is: (1). use the first stage to ensure the patient safety while exploring doses in the dose-finding procedure using continuous toxicity profiles (e.g., nTTP) from multiple toxicity types and multiple treatment cycles; (2). recruit additional

Table 2.3: Simulation results for scenario 11 to 14: $\mathbb{P}_{\text{alloc}}$, the percentage of dose allocation, $\mathbb{P}_{\text{recom}}$, the percentage of dose recommendation, where we recommend the lowest (safest) dose that is also efficacious, and $\mathbb{P}_{\text{effy}}$, the percentage of dose being efficacious and safe. mnTTP represents the mean nTTP score while pDLT refers to the probability of DLT event. In Scenario 11, all doses are too toxic, with flat cycle effect; In Scenario 12, only dose 1 is non-toxic, with flat cycle effect; Scenario 13/14 has an increasing/decreasing cycle effect, whose target toxicity dose at cycle 1 of the treatment is dose 5. Scenario 11-14 share a parabolic efficacy structure.

	Dose Levels						
	1	2	3	4	5	6	None
	Efficacy Structures						
Mean Efficacy	0.3	0.6	0.8	0.6	0.3	0.1	
	Scenario 11						
mnTTP(cycle 1 through 6)	0.359	0.409	0.433	0.439	0.452	0.465	
pDLT	0.446	0.513	0.558	0.559	0.588	0.613	
$\mathbb{P}_{\text{alloc}}$	88.0	12.0	0.0	0.0	0.0	0.0	
$\mathbb{P}_{\text{recom}}$	0.0	0.0	0.0	0.0	0.0	0.0	100.0
$\mathbb{P}_{\text{effy}}$	0.0	0.0	0.0	0.0	0.0	0.0	100.0
	Scenario 12						
mnTTP(cycle 1 through 6)	0.183	0.409	0.433	0.439	0.452	0.465	
pDLT	0.195	0.513	0.558	0.559	0.588	0.613	
$\mathbb{P}_{\text{alloc}}$	79.4	20.6	0.0	0.0	0.0	0.0	
$\mathbb{P}_{\text{recom}}$	94.6	0.1	0.0	0.0	0.0	0.0	5.3
$\mathbb{P}_{\text{effy}}$	94.6	0.1	0.0	0.0	0.0	0.0	5.3
	Scenario 13						
mnTTP(cycle 1)	0.046	0.054	0.108	0.183	0.280	0.359	
mnTTP(cycle 2)	0.054	0.063	0.124	0.206	0.303	0.381	
mnTTP(cycle 3)	0.063	0.073	0.142	0.230	0.326	0.404	
mnTTP(cycle 4)	0.073	0.085	0.161	0.255	0.349	0.426	
mnTTP(cycle 5)	0.085	0.097	0.181	0.280	0.373	0.448	
mnTTP(cycle 6)	0.097	0.110	0.203	0.307	0.396	0.469	
$\mathbb{P}_{\text{alloc}}$	17.0	20.4	22.7	22.3	13.8	3.8	
$\mathbb{P}_{\text{recom}}$	1.6	39.7	58.5	0.1	0.0	0.0	0.1
$\mathbb{P}_{\text{effy}}$	1.6	41.3	99.6	57.6	0.0	0.0	0.1
	Scenario 14						
mnTTP(cycle 1)	0.046	0.054	0.108	0.183	0.280	0.359	
mnTTP(cycle 2)	0.039	0.046	0.093	0.162	0.257	0.336	
mnTTP(cycle 3)	0.032	0.039	0.080	0.141	0.234	0.314	
mnTTP(cycle 4)	0.027	0.033	0.068	0.123	0.213	0.291	
mnTTP(cycle 5)	0.022	0.027	0.057	0.106	0.192	0.269	
mnTTP(cycle 6)	0.018	0.023	0.047	0.090	0.171	0.247	
$\mathbb{P}_{\text{alloc}}$	16.1	19.0	20.9	21.3	17.1	5.6	
$\mathbb{P}_{\text{recom}}$	2.8	39.8	57.1	0.1	0.1	0.0	0.1
$\mathbb{P}_{\text{effy}}$	2.8	42.5	99.4	62.3	0.8	0.0	0.1

patients in the second stage to explore the early-stage efficacy endpoint among safe doses; and (3). combine all data from all treated patients at the end of the study to find the lowest possible (safest) dose that is also efficacious. The proposed design can easily be generalized to any continuous / quasi-continuous toxicity endpoints, including TTB, TBS, and etc. To our best knowledge, this is the first design to combine an early-stage efficacy outcome and longitudinal toxicity endpoints over multiple treatment cycles in the dose-finding studies.

Simulation studies have been conducted to demonstrate that the proposed design consistently finds the safe doses with desired efficacy profiles in a wide range of scenarios. A Bayesian framework with MCMC is adopted because it naturally fits with the adaptive nature of Phase I dose-finding trials.

The design also allows for patient dropouts due to DLT, Therefore, missing data is present in both the longitudinal toxicities and the efficacy outcome. Around $20\% - 40\%$ of efficacy data is missing in our simulation studies depending on the scenarios. One exception is scenario 11 where all doses are too toxic and almost all efficacy data are missing since very few patients continue to cycle 3 of the treatment. In addition, our efficacy data generating distribution has quite a large amount of skewness / variation. Under these conditions, our joint modeling performs quite well, as shown by simulation results in Section 2.4.4.

The choices of c_1 and c_2 are picked based on the trial: if more toxicity is acceptable, then a higher upper bound of toxicity than the target toxicity can be chosen,

and p_1 and p_2 can consequently be picked by simulations. In simulations, we also tried $c_1 = c_2 = 0.36$, larger than the target toxicity 0.28, and p_1 and p_2 were set at 0.8. The maximally tolerated toxicity, therefore, was 0.36. The simulation result for scenario 1 showed a higher percentage of recommendations, 42%, for dose 6, whose mnTTP is 0.36 and thus an allowable (safe) dose in this setting. It has the largest mean efficacy of 0.9, among all doses. Another 52% of the recommendations correspond to the second best dose, dose 5, with a mean efficacy of 0.7. Around 94% of the trials, in total for scenario 1, recommend these two best doses. Across all scenarios, the design is more likely to define the target toxicity dose, with mnTPP 0.28, as the allowable (safe) dose, thus recommending this dose more frequently if it is the only efficacious dose. The selection of different weight matrices of the nTTP-based toxicity profiles and the selection of an appropriate efficacy endpoint can also have an impact on the dose-finding. Close collaborations with physicians and medical researchers are therefore required for this design, as with any other trials. Our method assumes that the same dose is given to a patient in each cycle of the treatment, thus it is an area of future research to allow dose modification across treatment cycles.

In summary, this chapter presents a new dose-finding method that accommodates an early efficacy endpoint and multiple toxicity types/grades over multiple treatment cycles in the dose-finding and dose recommendation for future studies. The simulations demonstrate that the proposed design performs well

in terms of identifying the optimal doses in various scenarios. An R package (**phase1RMD**) has been developed on CRAN to facilitate the use of the design (`https://cran.r-project.org/web/packages/phase1RMD/index.html`).

2.6 Appendix

2.6.1 Generation of Toxicity Data

In order to generate toxicity data, nTTP, for each patient, we need to define a matrix of probabilities, for each type of toxicity and each cycle of the treatment, of observing grades 0 to 4 toxicity, for K dose levels. For example, a possible such matrix, for cycle 1 of the treatment and renal toxicity, where $K = 6$, looks as below:

$$\mathbb{P}_{\text{renal}} = \begin{pmatrix} 0.823 & 0.152 & 0.022 & 0.002 & 0.001 \\ 0.791 & 0.172 & 0.032 & 0.004 & 0.001 \\ 0.758 & 0.180 & 0.043 & 0.010 & 0.009 \\ 0.685 & 0.190 & 0.068 & 0.044 & 0.013 \\ 0.662 & 0.200 & 0.078 & 0.046 & 0.014 \\ 0.605 & 0.223 & 0.082 & 0.070 & 0.020 \end{pmatrix},$$

where grades from 0 to 4 define the columns of the matrix while the dose levels from 1 to 6 define the rows. Such matrices for all three types of toxicities and all

treatment cycles define a toxicity structure, where the mnTTP (mean nTTP) for each dose level and cycle can be derived. When all such matrices are known, we can compute, for each dose level and cycle, the probabilities (weights) of each of the 5^3 combinations of the TPs (toxicity profile). For each combination, we can also compute its nTTP score from the clinical weight matrix. The weighted sum of nTTP for all combinations is thus the mnTTP associated with the dose level and the cycle. The pDLT (probability of DLT) can be computed similarly.

These probability matrices are used to generate nTTP for each patient, given the dose allocated and the cycle of the treatment. For example, suppose a patient is assigned dose 1 at cycle 1 of the treatment. For the renal toxicity, we look at the first row of matrix $\mathbb{P}_{\text{renal}}$, $(0.823, 0.152, 0.022, 0.002, 0.001)$, which corresponds to the probabilities of observing renal toxicity grades 0 to 4 respectively at dose 1. We thus sample a renal toxicity grade for this patient, according to this probability distribution. We repeat this sampling for the other two types of toxicities and in the end, we have a combination of the TP for the patient, for example, grade 1 renal associated with grade 1 neurological and grade 3 hematological toxicities. The nTTP is thus given by the clinical weight matrix, such that $\text{nTTP} = \sqrt{0.5^2 + 0.5^2 + 0.5^2}/2.5 = 0.35$. We also assess if the patient experiences a DLT by its definition, and if a patient has a DLT, he/she will drop out for the subsequent cycles, if any.

2.6.2 Generation of Efficacy Outcome

Figure 2.1b to 2.1f show the distribution of efficacy outcome for dose-efficacy pattern 1 to 5 across each dose level. Pattern 1 (Figure 2.1b) corresponds to an increasing relationship between efficacy and dose, as reflected by the red solid line in Figure 2.1a, "Dose-Efficacy Patterns", where the mean of efficacy outcome increases as dose level goes up. Pattern 4 (Figure 2.1e), the blue dashed line in Figure 2.1a, represents a decreasing relationship, opposite to the pattern 1. Pattern 2 (Figure 2.1c) , a flat line, indicates a flat relationship where the mean of efficacy outcome does not change with dose. Pattern 3 (Figure 2.1d) demonstrates a situation where the mean of efficacy outcome increases and then reaches a plateau as dose level increases, corresponding to the green dashed line in Figure 2.1a and let us call this plateau relationship. The last pattern, pattern 5 (Figure 2.1f), shows a parabolic relationship, as seen from the purple dotted line in Figure 2.1a that the mean of efficacy outcome first increases and then decreases as dose level varies from low to high.

It is clearly seen from dose-efficacy pattern 1 to 5 that efficacy data distributions have large variation and/or skewness.

Chapter 3

Phase II Studies: An Adaptive Enrichment Design to Determine the Population to Enroll in Phase III Trials

SUMMARY.[1]

We propose and evaluate a two-stage, Phase II, adaptive clinical trial design. Its goal is to determine whether future Phase III (confirmatory) trials should be conducted, and if so, which population should be enrolled. The population selected

[1]This Chapter 3 is adapted from the working paper "**Yu Du**, Gary L. Rosner, Michael Rosenblum. *Phase II Adaptive Enrichment Design to Determine the Population to Enroll in Phase III Trials, by Selecting Thresholds for Baseline Disease Severity.*"

for Phase III enrollment is defined in terms of a disease severity score measured at baseline. We optimize the Phase II trial design and analysis in a decision theory framework. Our utility function represents a combination of the cost of conducting future Phase III trials and, if the Phase III trials are successful, the improved health of the future population minus the cost of treatment. Given such a utility function and a discrete prior distribution on the conditional treatment effect, we compute an approximate Bayes optimal adaptive design. The resulting design is compared to simpler designs in simulation studies. We also apply the designs to resampled data from a completed, Phase II trial evaluating a new surgical intervention for stroke.

3.1 Background

A new treatment may be effective only for a subset of the overall study population. This poses important challenges for designing a randomized clinical trial to evaluate such a treatment. On the one hand, ignoring participant heterogeneity and enrolling the overall population can lead to a dilution of the treatment effect, and may create an unethical situation where participants who do not benefit from the treatment are treated. On the other hand, enrolling only a narrow proportion of the overall population would not answer the question of whether the treatment benefits the larger population; furthermore, such restrictive enrollment could lead

to slower recruitment and longer trial duration.

Our work is motivated by a multicenter, randomized trial where the new surgical intervention called Minimally-Invasive Surgery Plus rt-PA for Intracerebral Hemorrhage Evacuation (MISTIE) was compared to standard medical care (Hanley et al., 2016). An important baseline (i.e., pre-randomization) characteristic is intracerebral hemorrhage (ICH) volume, which is one measure of disease severity. Based on their understanding of brain hemorrhage, the clinical investigators conjectured that the new treatment may have different effects depending on a participant's pre-randomization ICH volume. We compare different types of Phase II, randomized trial designs whose goal is to inform a future recommendation about which (if any) range of ICH volume should be used as the enrollment criterion in future Phase III, confirmatory trials. For computational reasons, we discretize ICH volume by partitioning the range of its possible values; in general, if one has a continuous-valued baseline score, our approach can be applied to a discrete version of it.

We consider two-stage, adaptive enrichment designs for the Phase II randomized trial. Adaptive enrichment designs involve preplanned rules for modifying enrollment criteria based on data accrued in an ongoing trial (Wang et al., 2009). We hypothesized that adaptive enrichment in the Phase II trial might be useful for making an optimal recommendation for the population (if any) to enroll in future Phase III confirmatory trials. In such a Phase II design, information from the

first stage of the Phase II trial can be used to target whom to enroll in the second stage of the Phase II trial. For example, if early Phase II data indicated a treatment benefit only for those with high baseline scores, participants near the boundary of such scores could be oversampled in the second stage of the Phase II trial; this may lead to improved information for recommending which population to enroll in Phase III. We investigate whether such an adaptive feature adds value (in terms of expected utility, defined below) or not, compared to a non-adaptive Phase II design.

Related work aimed at determining the population who benefits from a treatment includes, e.g., Jiang et al. (2007), Zhou et al. (2008), Barker et al. (2009), Freidlin et al. (2010), Lee et al. (2010), Kim et al. (2011), Cai et al. (2011), Lai et al. (2014), Xu et al. (2014), Ohwada and Morita (2016), Spencer et al. (2016). Our approach differs from these in that we explicitly define the performance goal of our Phase II design (the objective function) in a decision theory framework and compute an approximate Bayes optimal adaptive enrichment design over a class of designs defined in Section 3.2.

We next discuss related work that uses a decision theory approach for optimizing trial designs. Colton (1963, 1965) aim to select the best of two treatments; Banerjee and Tsiatis (2006) aim to minimize expected sample size; Cheng and Berry (2007) aim to maximize the expected number of effectively treated participants in the trial; Hampson and Jennison (2015) aim to optimize power to detect the best

of multiple treatments. Each of the aforementioned references involves a single population. In contrast, we consider multiple populations defined by a baseline score, and aim to determine which population (if any) to recommend for a pair of future Phase III trials. Our designs also differ from the above related work in terms of the types of designs we consider, i.e., Phase II designs that can adapt the population enrolled (called adaptive enrichment). The work of Graf et al. (2015), Krisam and Kieser (2015), Götte et al. (2015), and Rosenblum et al. (2016) involves adaptive enrichment using a binary biomarker that divides participants into two subpopulations. In contrast, by allowing the population selected for Phase III to be an interval of the baseline score, we allow a larger number of possible subpopulations.

In Section 3.2, we define the class of trial designs that we optimize over. The trial design optimization problem and our approach to solving it are given in Section 3.3. In Section 3.4, we use simulations to compare the performance of an optimized, two-stage, adaptive design versus an optimized (non-adaptive) one-stage design. The main result is that the two-stage, adaptive design did not improve expected utility, where the utility function measures the quality of the recommendation for whom to enroll in future Phase III trials; however, as shown in Section 3.4.4, the two-stage, adaptive design leads to fewer participants assigned to a non-efficacious or harmful treatment during Phase II. We apply the designs to resampled data from a completed, Phase II trial (MISTIE II) evaluating a new sur-

gical intervention for stroke, in Section 3.5. In Section 3.6, we present areas for future research.

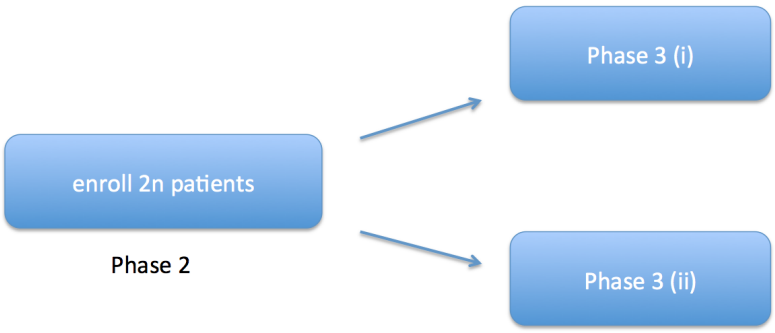
3.2 Data Generating Process and Phase II Trial Designs

3.2.1 Overview of Fixed and Adaptive Designs

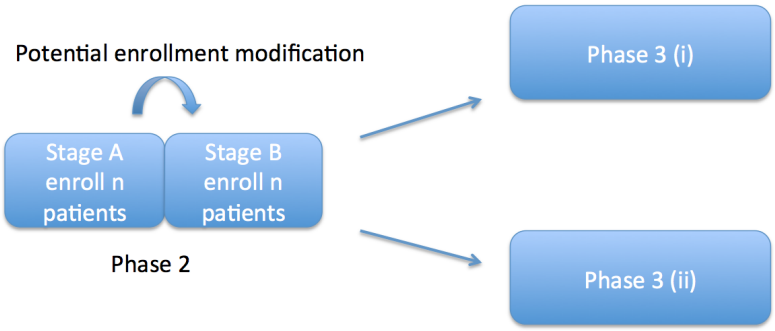
We focus on Phase II randomized trials comparing a new treatment to standard of care (control) with a 1:1 randomization ratio. We compare two-stage, adaptive enrichment designs to one-stage, fixed designs. Both designs have the same total sample size (denoted as $2n$), and both have the goal of making an optimal recommendation for the population (if any) to enroll in two, future Phase III trials. Two, future Phase III trials are considered since this is what the U.S. Food and Drug Administration typically requires for approval of a new drug; the Phase III trials are assumed to be fixed (non-adaptive) and have predefined sample size N.

Figure 3.1 illustrates two types of Phase II designs: the one-stage, fixed design and the two-stage, adaptive enrichment design, referred to as the fixed design and adaptive design, respectively. In the fixed design, shown in Figure 3.1a, $2n$ participants are enrolled from the overall population. A preplanned rule at the end of the trial determines the population to enroll in two future Phase III trials. The

population is defined by a range of the baseline score.



(a) Fixed Design, $2n$ participants are enrolled in Phase II from the overall population. At the end, a recommendation of the population (if any) to enroll in two, future Phase III trials is made.



(b) Adaptive Design, n participants are enrolled in each of Stages A and B in Phase II with the Stage B enrollment criterion depending on the results of Stage A. After Stage B, a recommendation of the population (if any) to enroll in two, future Phase III trials is made.

Figure 3.1: Phase II fixed design (top-left) and Phase II adaptive design (bottom-left), each of which may be followed by two Phase III trials.

The adaptive design, shown in Figure 3.1b, has two stages. In Stage A, n participants are recruited from the overall population. At the end of Stage A, a decision is made as to which population to enroll in Stage B, in terms of the baseline score. This allows an enrollment modification for Stage B (which always has n

participants), with the goal of providing more targeted information to help in the population selection at the end of Stage B for the future Phase III trials.

For both the fixed and adaptive Phase II designs, the data from the $2n$ participants are ultimately used to select a range of the baseline score to set as the enrollment criterion for 2 future Phase III trials. If the range selected is the empty set, this corresponds to not conducting any Phase III trials; this option is important since one of the main goals of Phase II is to weed out useless treatments.

For the Phase II adaptive design, the only feature that is adapted is the Stage B enrollment criterion. The decisions involved in the Phase II adaptive designs:

(i) After collecting the Stage A data, what will be the enrollment criterion will be for Stage B?

(ii) After Stage B data has been collected, what will be the enrollment criterion for the future Phase III trials?

3.2.2 Data Collected on Each Participant

Let R denote the continuous-valued baseline score, e.g., ICH volume in the MISTIE trial example. We assume the population distribution of R is uniform on the interval $(0, 1)$, which can be achieved by a quantile transformation of any continuous variable. In order to make our computations feasible, we discretize R by partitioning $(0, 1)$ into M consecutive, equal length intervals. Let $\tilde{R} = \lceil R \times M \rceil$

denote the corresponding interval number, which has values in $\{1, \ldots, M\}$. We refer to $\tilde{R}$ as the discrete score.

Let T denote the treatment arm indicator, where $T = 1$ means assignment to treatment and $T = 0$ means assignment to control. Let $Y \in \mathbb{R}$ denote the primary outcome, which we assume to be measured on each participant relatively soon after her/his enrollment. Let $\Delta : (0, 1) \to \mathbb{R}$ denote the conditional treatment effect function $\Delta(r) = E(Y|T = 1, R = r) - E(Y|T = 0, R = r)$ given the continuous-valued baseline score $r \in (0, 1)$. Define the discrete analog of Δ, which represents the average treatment effect for stratum $\tilde{R} = \tilde{r}$, as:

$$\tilde{\Delta}(\tilde{r}) = E(Y|T = 1, \tilde{R} = \tilde{r}) - E(Y|T = 0, \tilde{R} = \tilde{r}) = M \int_{(\tilde{r}-1)/M}^{\tilde{r}/M} \Delta(r)dr, \qquad (3.1)$$

for any $\tilde{r} \in \{1, \ldots, M\}$.

Conditional on the discrete baseline score and treatment indicator, the primary outcome Y is assumed to be a random, independent draw from a normal distribution with common variance σ^2. That is, we assume the conditional distribution of $Y|T = t, \tilde{R} = \tilde{r}$ is Normal$(\mu_t(\tilde{r}), \sigma^2)$, for unknown mean functions $\mu_0(\tilde{r}), \mu_1(\tilde{r})$ satisfying $\mu_1(\tilde{r}) - \mu_0(\tilde{r}) = \tilde{\Delta}(\tilde{r})$ and common, conditional variance σ^2. For simplicity, we assume $\mu_1(\tilde{r}) = \tilde{\Delta}(\tilde{r})/2$ and $\mu_0(\tilde{r}) = -\tilde{\Delta}(\tilde{r})/2$.

The data vector contributed by each participant i, used as input to the decision rules in the trial design, is denoted $V_i = (\tilde{R}_i, T_i, Y_i)$. Let $X^{(A)}$ and $X^{(B)}$ denote

the sets of data vectors V_i collected during Stage A and Stage B, respectively. Let $X = (X^{(A)}, X^{(B)})$ denote the entire data set at the end of Stage B.

3.2.3 Definition of Phase II Adaptive Designs

Let $\mathcal{E}^{(A)}$ and $\mathcal{E}^{(B)}$ denote the allowed enrollment choices at the end of Stage A and Stage B, respectively. The action sets $\mathcal{E}^{(A)}, \mathcal{E}^{(B)}$ each consist of a prespecified, finite set of intervals (r_l, r_u) with endpoints $r_l, r_u \in \{0, 1/M, 2/M, \ldots, 1\}$ and $r_l \leq r_u$; each interval represents a range of the baseline score. If the interval $(r_l, r_u) \in \mathcal{E}^{(A)}$ is selected at the end of Stage A, it means that n participants will be enrolled during Stage B using inclusion criterion $R \in (r_l, r_u)$. If the interval $(r_l, r_u) \in \mathcal{E}^{(B)}$ is selected at the end of Stage B, it means that the two, Phase III trials will enroll N participants using inclusion criterion $R \in (r_l, r_u)$. We assume that $\mathcal{E}^{(A)}$ contains the full interval $(0, 1)$ and that $\mathcal{E}^{(B)} = \mathcal{E}^{(A)} \cup \{\emptyset\}$, where the empty set $\emptyset$ represents not conducting any Phase III trials. By convention, we represent the empty set by an interval (r_l, r_u) with $r_l = r_u$.

An adaptive design is defined as a pair of decision rules $(d^{(A)}, d^{(B)})$, where $d^{(A)}$ is a map from the Stage A data $X^{(A)}$ to the set of enrollment choices $\mathcal{E}^{(A)}$, and $d^{(B)}$ is a map from the cumulative data X to $\mathcal{E}^{(B)}$. We assume these maps are measurable, which is generally required for adaptive designs (Liu et al., 2002). Whenever maxima are taken over pairs of decision rules $(d^{(A)}, d^{(B)})$, we assume that this is over all possible pairs of such measurable maps.

After collecting the Stage A data $X^{(A)}$, the prespecified decision rule $d^{(A)}$ is applied to select the action from $\mathcal{E}^{(A)}$; this determines the population to be enrolled during Stage B. At the end of Stage B, the cumulative data $X = (X^{(A)}, X^{(B)})$ has been collected, and the prespecified decision rule $d^{(B)}$ is applied to determine the action in $\mathcal{E}^{(B)}$; this action represents the enrollment criterion for the two, future Phase III confirmatory trials. The fixed design is a special case of the adaptive design with $d^{(A)}$ always mapping to the action $(0, 1)$.

3.2.4 Data Generating Process for Adaptive Phase II Design

The data generating process for a trial conducted using an adaptive design $(d^{(A)}, d^{(B)})$ is summarized in Figure 3.2. First, the treatment effect function Δ is drawn from the prior distribution π; this determines the conditional mean functions μ_0, μ_1 as defined in Section 3.2.2. Next, Stage A data from n participants are generated, where each participant has baseline score R drawn from the uniform distribution over $(0, 1)$ corresponding to the full range of the severity score; this determines her/his discrete baseline score $\tilde{R}$. Treatment indicator T is assigned independent of the baseline score with probability $1/2$ of being 0 or 1. Next, the outcome is drawn according to the conditional distribution $Y|T = t, \tilde{R} = \tilde{r}$, which was assumed to be Normal$(\mu_t(\tilde{r}), \sigma^2)$.

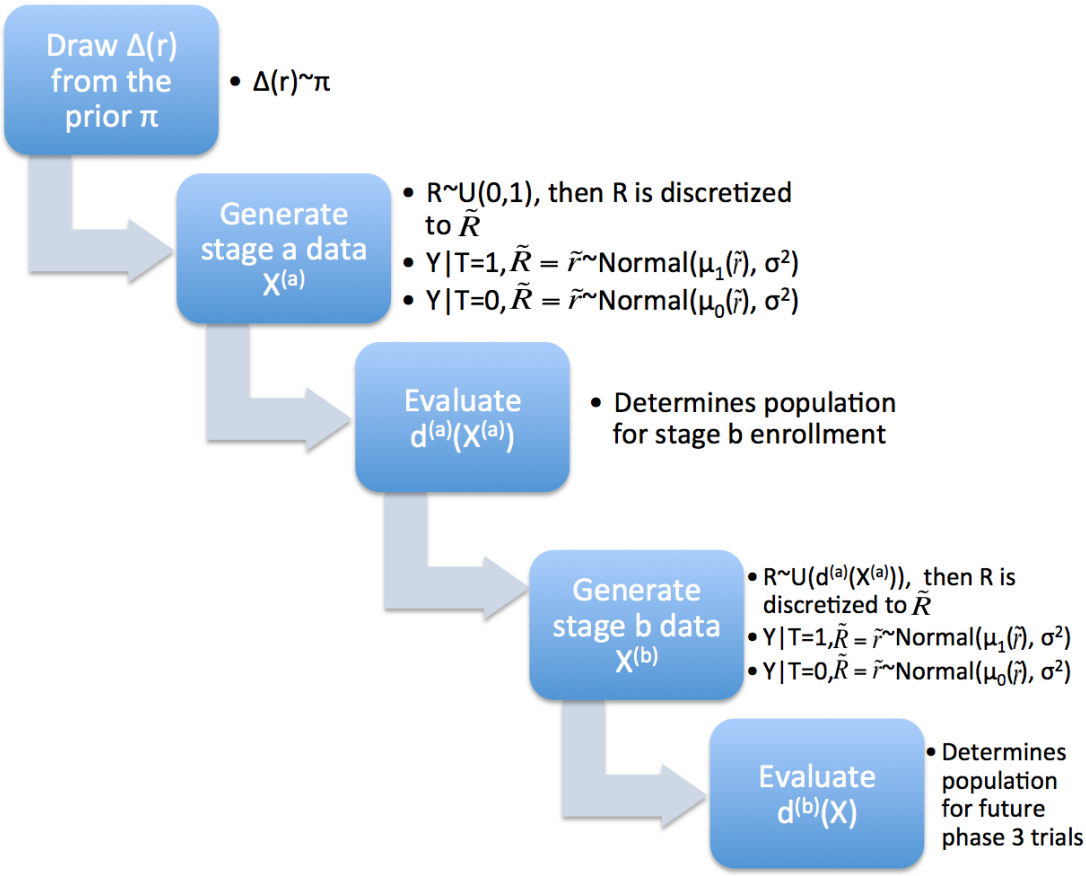


Figure 3.2: Data generating process for Phase II adaptive design using decision rule $(d^{(A)}, d^{(B)})$.

In the third step, the decision function $d^{(A)} \in \mathcal{D}^{(A)}$ maps the Stage A data $X^{(A)}$ to an action in $\mathcal{E}^{(A)}$. This action represents the population enrolled during Stage B. Each Stage B participant has baseline score R drawn uniformly from the interval $d^{(A)}(X^{(A)})$. The treatment assignments and outcomes given $\tilde{R}$ are drawn analogously as in Stage A. At the end of Stage B, given the cumulative data X, the decision function $d^{(B)}$ determines the population to enroll in the Phase III trials, including the option to conduct no future trials.

3.2.5 Prior Distribution on Conditional Treatment Effect Function

We assume a prior π on the conditional treatment effect function Δ. In our simulations, π is set to be a finite set of point masses on the functions $\Delta = \delta_1, \ldots, \delta_6$, shown in Figure 3.3. We refer to each δ_k as a possible state of nature.

In Figure 3.3, the function δ_1 represents the conditional treatment effect being 1 for all values of the baseline score R in $(0, 1)$. The function δ_2 represents the treatment only benefiting the population with baseline scores greater than $1/2$, while the function δ_4 represents the treatment benefiting the population whose baseline scores are between $1/4$ and $3/4$. Under the conditional treatment effect function δ_5, the treatment benefit is -1, i.e., the treatment is harmful, for all values of the baseline score. Under δ_6, there is zero treatment benefit for every baseline score value; this represents the global null hypothesis of no treatment effect for any stratum of the baseline score. The discrete versions of $\delta_1, \ldots, \delta_6$, based on applying (3.1) to each, are given in Table 3.5 of the Appendix.

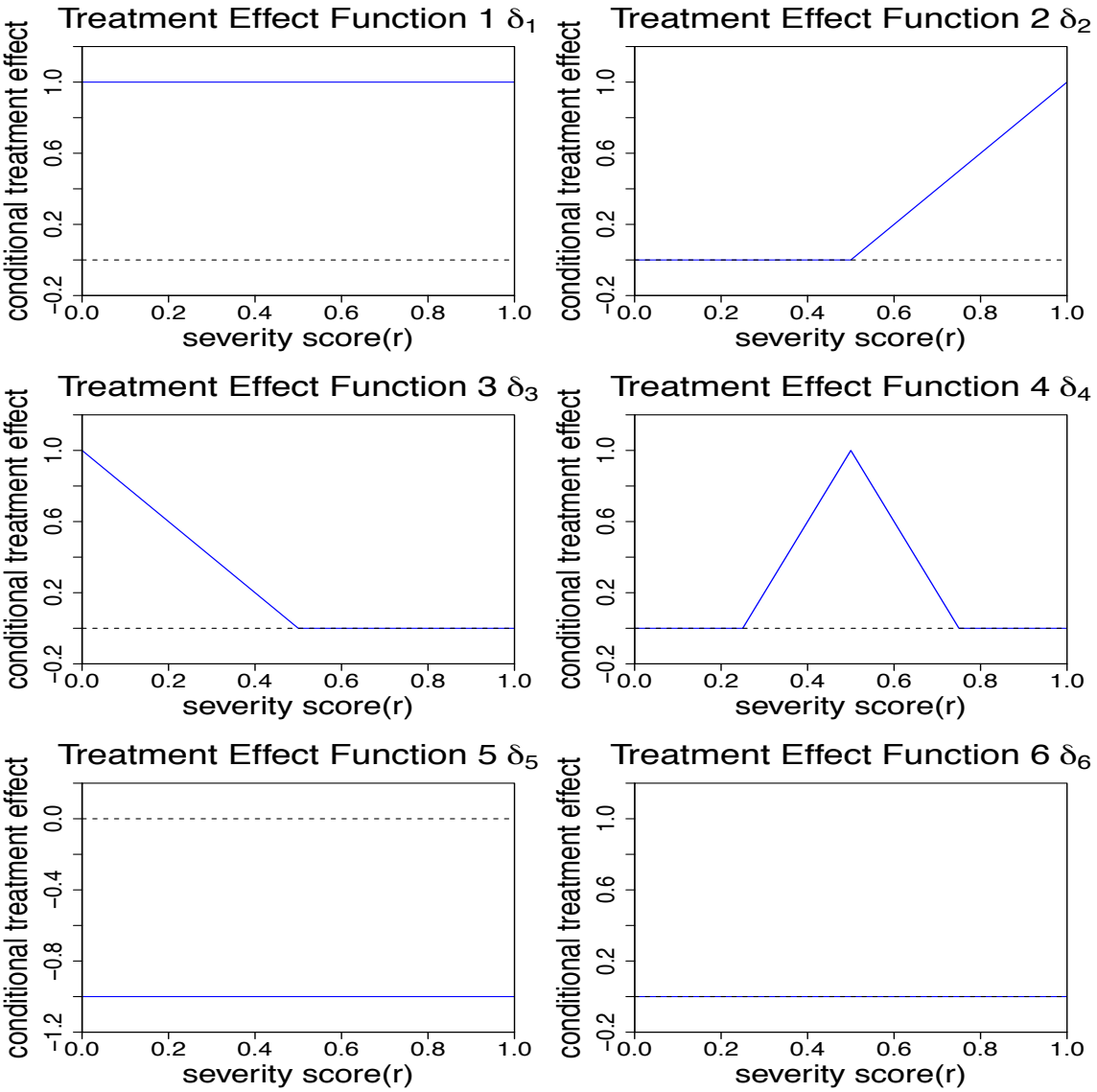


Figure 3.3: Six states of nature $\delta_1, \ldots, \delta_K$. The solid line represents the treatment effect function δ_k, while the dashed line is a horizontal zero line as a reference.

3.3 Optimization Problem and Algorithm to Solve It

3.3.1 Utility Function

Our utility function focuses on what happens as a consequence of the Phase II trial. It represents a combination of the cost of conducting future Phase III trials (if they are recommended after Phase II) and the improved health of the future population minus the treatment's cost (if the Phase III trials succeed). The utility function $U(r_l, r_u; \Delta)$ takes the function Δ and an action $(r_l, r_u) \in \mathcal{E}^{(B)}$ (which represents the interval recommended for Phase III enrollment) as inputs, and is defined as follows: if no Phase III trials are recommended (which occurs when $r_l = r_u$) then $U(r_l, r_u; \Delta) = 0$, and otherwise $U(r_l, r_u; \Delta)$ equals

$$P(\text{2 Phase III Trials Succeed}|\Delta, r_l, r_u) \times \left\{ \overbrace{\int_{r_l}^{r_u} \Delta(r)dr}^{\text{Treatment Benefit}} - \overbrace{c(r_u - r_l)}^{\text{Treatment Cost}} \right\} - \overbrace{\lambda/(r_u - r_l)}^{\text{Phase III Cost}}, \tag{3.2}$$

where $P(\text{2 Phase III Trials Succeed}|\Delta, r_l, r_u)$ is the conditional probability, defined below, that both Phase III trials enrolling those with baseline scores in (r_l, r_u) succeed given Δ.

The components in curly braces in (3.2) represent the population health and

treatment cost due to the treatment being approved (after 2 successful Phase III trials) for use in the population with baseline scores in the range (r_l, r_u). The term $\int_{r_l}^{r_u} \Delta(r)dr$ represents the health benefit that would result from the population with baseline score in the interval (r_l, r_u) using the treatment instead of the control. This term is maximized if the interval (r_l, r_u) contains precisely those who benefit from treatment (and possibly also those who have no effect from treatment).

The second term in curly braces $c(r_u - r_l)$ represents health system costs to the future population if both Phase III trials succeed leading to the drug being approved and used by the recommended population (r_l, r_u). Here, c is a constant, and we assume that the health system cost is proportional to the size of the population with baseline score in (r_l, r_u). We use the term "cost" in the general sense that captures negative impacts of administering the treatment to a future population. For example, cost may reflect monetary cost and negative impacts in terms of frequency and severity of side effects caused by treatment. If the cost $c > 0$, then there is a penalty for approving the treatment for any stratum $r \in (0, 1)$, especially for those whose treatment effect $\Delta(r)$ is less than c; in particular, it penalizes for approving the treatment for strata that have zero treatment effect (unlike the first term in curly braces).

The component $P(\text{2 Phase III Trials Succeed}|\Delta, r_l, r_u)$ is the probability of success of two, future Phase III trials enrolling from the population with baseline score in the interval (r_l, r_u). Assuming that the two Phase III trials are independent given

(r_l, r_u), this probability is the squared power of a fixed (non-adaptive) design for the Phase III trial where the average treatment effect is $(r_u - r_l)^{-1} \int_{r_l}^{r_u} \Delta(r)dr$, i.e., the average height of the treatment effect curve Δ over the interval (r_l, r_u). The prespecified sample size N for each Phase III trial is determined by the minimum, clinically meaningful benefit Δ_{min}, type I error, type II error, and outcome variance σ^2.

The last term of the utility function, $\lambda/(r_u - r_l)$, is proportional to the duration of each Phase III trial enrolling participants with baseline scores in (r_l, r_u), where we made the approximation that the enrollment rate is proportional to the size of the recommended population to be enrolled. This could apply, for example, to the MISTIE trial where ICH volume is only determined after recruitment based on neuroimaging, and so excluding more participants would require longer recruitment to reach the total sample size N.

Though Type I error control in the Phase II trial is not our focus, we do evaluate design performance when there is no treatment effect to determine how often future Phase III trials are (erroneously) recommended.

3.3.2 Optimization Problem

The goal is to compute the pair of optimal decision rules at the end of Stage A and at the end of Stage B in Phase II, denoted by $d_{opt}^{(A)}$ and $d_{opt}^{(B)}$, respectively. These are the rules that maximize expected utility, where the expectation is with respect

to the distribution of (Δ, X) induced by the prior π on Δ. Define

$$(d_{opt}^{(A)}, d_{opt}^{(B)}) \;=\; \underset{d^{(A)}, d^{(B)}}{\operatorname{argmax}} \mathbb{E}\left[U\left\{d^{(B)}\left(X^{(A)}, X^{(B)}[d^{(A)}\{X^{(A)}\}]\right); \Delta\right\}\right], \tag{3.3}$$

where the maximum is taken over all pairs $(d^{(A)}, d^{(B)})$ of decision rules as defined in Section 3.2.3. Throughout, expectation $\mathbb{E}$ and probability P are with respect to the prior π and a generic Phase II adaptive design $(d^{(A)}, d^{(B)})$ unless indicated otherwise. As described in Section 3.3.1, the utility U depends on the action $d^{(B)}(X)$ taken at the end of Stage B (i.e., who to enroll in the future Phase III trials) and the conditional treatment effect function Δ. The action taken at the end of Stage B depends on the data from Stages A and B, i.e., $X = (X^{(A)}, X^{(B)})$. We write $X^{(B)} = X^{(B)}[d^{(A)}\{X^{(A)}\}]$ in the display above to make explicit that the Stage B data generating distribution depends on the decision $d^{(A)}\{X^{(A)}\}$ at the end of Stage A regarding the population to be enrolled during Stage B. This highlights the sequential nature of the decision problem.

3.3.3 Algorithm to Solve the Optimization Problem

Banerjee and Tsiatis (2006), Cheng and Berry (2007) and Hampson and Jennison (2015) used backward induction to provide closed form solutions to their optimal adaptive designs under a decision-theoretic framework. Since this is not possible in our problem, we instead use backward induction with Monte-Carlo forward

simulation. Such backward induction has been implemented to solve problems in different contexts by, e.g., Carlin et al. (1998), Brockwell and Kadane (2003), Rossell et al. (2006), Ding et al. (2008).

Let $n^{(j)}(\tilde{r})$ denote the cumulative sample size per arm for participants with baseline score $\tilde{R} = \tilde{r}$ enrolled during or before Stage $j \in \{A, B\}$ of the Phase II trial, for each $\tilde{r} \in \{1, \ldots, M\}$. We assume that in Stage A, n/M are enrolled from each baseline stratum $\tilde{r} \in \{1, \ldots, M\}$. For Stage B of the adaptive design, we assume an equal number are enrolled from each stratum $\tilde{r}$ contained in the selected population $d^{(A)}(X^{(A)})$, such that a total of n are enrolled in that stage. We assume that within each stage and enrolled stratum of the baseline score, an equal number are assigned to treatment and control; this can be accomplished (approximately) by stratified block randomization.

Define the sample mean difference between arms of the primary outcome using cumulative data through the end of Stage j for participants with $\tilde{R} = \tilde{r}$ as

$$\widehat{\Delta}^{(j)}(\tilde{r}) = \frac{1}{n^{(j)}(\tilde{r})} \left\{ \sum_i 1(T_i = 1, \tilde{R}_i = \tilde{r}) Y_i - \sum_i 1(T_i = 0, \tilde{R}_i = \tilde{r}) Y_i \right\},$$

where the summations are over the participants i with outcomes observed at or before the end of Stage j, and $1(S)$ is the indicator variable taking value 1 if S is true and 0 otherwise. The statistic $\widehat{\Delta}^{(j)}(\tilde{r})$ is an estimator of $\Delta(\tilde{r})$, the average treatment effect in stratum $\tilde{r}$. Let $\widehat{\Delta}^{(j)}$ denote the vector of cumulative sample

mean differences for each category of baseline score $\tilde{r} \in \{1, \ldots, M\}$ using all data up through the end of Stage $j \in \{A, B\}$.

We next define minimal sufficient statistics $\tilde{S}^{(A)}, \tilde{S}^{(B)}$ for Δ based on the data $X^{(A)}$ available at the end of Stage A and the data X available at the end of Stage B, respectively. Stage A is equivalent to a fixed design, and we define $\tilde{S}^{(A)} = \widehat{\Delta}^{(A)}$, i.e., the sample mean differences within each stratum $\tilde{r} \in \{1, \ldots, M\}$. Since Stage B involves a potential enrollment modification that is determined by the decision $d^{(A)}(X^{(A)})$, we define $\tilde{S}^{(B)} = (d^{(A)}(X^{(A)}), \widehat{\Delta}^{(B)})$.

Backward induction starts from the end of Stage B, where we have collected the overall data $X = (X^{(A)}, X^{(B)})$. The first step is to maximize the conditional expected utility over all possible Stage B decision rules given X. It follows from (3.3) that

$$d^{(B)}_{opt}(X) = \underset{d^{(B)}}{\operatorname{argmax}}\, \mathbb{E}[U\{d^{(B)}(X); \Delta\}|X], \tag{3.4}$$

where the expectation is with respect to the distribution of Δ given X.

We assume the prior distribution π on Δ consists of K point masses $\delta_1, \ldots, \delta_k$, each representing a conditional treatment effect function. For any candidate action $(r_l, r_u) \in \mathcal{E}^{(B)}$, Monte Carlo simulation is used where we draw posterior samples of Δ given the data X in order to approximate the conditional expected utility if this action is followed. i.e., $\mathbb{E}[U(r_l, r_u; \Delta)|X]$. The posterior distribution $P(\Delta \mid X)$

depends on the data X only through the sufficient statistics $\tilde{S}^{(B)}$. Given the data X, we compute the corresponding sufficient statistics $\tilde{S}^{(B)}$ and then draw posterior samples of Δ from $P(\Delta \mid \tilde{S}^{(B)})$ as described below.

Consider any Stage A decision rule $d^{(A)}$ and data vector $X = (X^{(A)}, X^{(B)})$ that can be generated under the Phase II adaptive design using Stage A decision rule $d^{(A)}$. Let $E = d^{(A)}(X^{(A)})$ denote the decision at the end of Stage A; without loss of generality, we assume $d^{(A)}$ depends on the data $X^{(A)}$ only through the sufficient statistic $\hat{\Delta}^{(A)}$. We prove the following in the Appendix, where $\propto$ represents proportionality with respect to functions of the data X:

$$P(\Delta = \delta_k \mid X) \propto P^E(\widehat{\Delta}^{(B)} = \widehat{\Delta}^{(B)}(X) \mid \Delta = \delta_k) P(\Delta = \delta_k), \tag{3.5}$$

where P^E denotes the probability density function of X under the deterministic (non-adaptive) Stage A decision rule that always enrolls population $E \in \mathcal{E}^{(A)}$ during Stage B regardless of the Stage A data. The above display reduces (up to a proportionality constant) the problem of computing $P(\Delta = \delta_k \mid X)$ to the following simpler problem: computing the conditional probability density that the cumulative sample mean differences (at the end of Stage B) equal $\widehat{\Delta}^{(B)}(X)$ given Δ under the non-adaptive Stage A decision rule that always enrolls population $E = d^{(A)}(X^{(A)})$ during Stage B.

We describe how to compute $d^{(B)}_{opt}(X)$ given X. Under the probability density

P^E, the statistic $\widehat{\Delta}^{(B)}$ conditional on $\Delta = \delta_k$ has a multivariate normal distribution with mean vector determined by integrating δ_k over each interval $((m-1)/M, m/M) : m = 1, \ldots, M$ using the formula on the right side of (3.1) and covariance matrix $\mathbf{\Sigma}$ the diagonal matrix with zeros off the main diagonal and $\mathbf{\Sigma}_{mm} = 2\sigma^2/n^{(B)}(m)$ for each $m = 1, 2, \ldots, M$. Given the observed $\widehat{\Delta}^{(B)}$, for each $k = 1, \ldots, K$ we compute the density $P^E(\widehat{\Delta}^{(B)} = \widehat{\Delta}^{(B)}(X) \mid \Delta = \delta_k)$ from this multivariate normal distribution and multiply by the prior $P(\Delta = \delta_k)$ to obtain the right side of (3.5), which was shown above to be proportional to the posterior probability $P(\Delta = \delta_k \mid X)$. For each interval (r_l, r_u) in the action space $\mathcal{E}^{(B)}$, we repeatedly draw from this posterior distribution to approximate the posterior expected utility $\sum_{k \leq K} U(r_l, r_u; \delta_k) P(\Delta = \delta_k \mid X)$, which is $\mathbb{E}[U\{d^{(B)}(X); \Delta\}|X]$ under the decision $d^{(B)}(X) = (r_l, r_u)$. We set $d^{(B)}_{opt}(X)$ to be the interval (r_l, r_u) in $\mathcal{E}^{(B)}$ that gives the maximum posterior expected utility.

The derivation of $d^{(A)}_{opt}$ is analogous to that of $d^{(B)}_{opt}$, and is described in the Appendix. All of the above computations were implemented in R (R Core Team, 2015).

3.4 Simulation Study with Six Possible Treatment Effect Curves

3.4.1 Simulation Setup

We implement the optimization algorithm from the previous section in a simulation study comparing the performance of the fixed versus adaptive Phase II design. The prior π on Δ is a discrete distribution on the conditional treatment effect curves $\{\delta_k\}_{k=1}^6$ in Figure 3.3 that places 50% weight on δ_6 (the global null hypothesis) and equal weights on $\delta_1, \ldots, \delta_5$. We allocate greater weight to δ_6 in order to reflect the realistic possibility that an experimental treatment has no effect at all. Intuitively, the impact of having greater weight on δ_6 is that the corresponding optimal design will be more conservative in initiating Phase III follow-up trials, thereby saving resources at the cost of lowering the chance of successful Phase III trials.

The prior π is one possible choice; our general approach can be applied to an arbitrary set of weights and any finite set of conditional treatment effect functions, i.e., applied to any prior consisting of a finite set of point masses. The choice of prior would ideally be chosen based on prior scientific knowledge and data from earlier studies. Here we demonstrate a relatively simple case as a proof of concept.

For a given state of nature Δ, *the population that benefits from treatment* is defined

as $\{r \in (0,1) : \Delta(r) > 0\}$. For example, under $\Delta = \delta_2$, the treatment benefits precisely those with $r \in (0.5, 1)$. Depending on which of $\delta_k, k = 1, \ldots, 6$, is the true state of nature, the population who benefit from treatment is an interval of the baseline score r in the set

$$\mathcal{E}^{(B)} = \{(0,1), (0.5,1), (0,0.5), (0.25,0.75), \emptyset\}. \tag{3.6}$$

Let

$$\mathcal{E}^{(A)} = \{(0,1), (0.5,1), (0,0.5), (0.25,0.75)\}.$$

We use the above sets $\mathcal{E}^{(A)}, \mathcal{E}^{(B)}$ as the action sets for our decision problem, representing the possible choices for whom to enroll next based on the data at the end of Stages A and B, respectively.

We define $\tilde{R}$ to have $M = 4$ levels corresponding to the baseline score being in the intervals $(0, 0.25), (0.25, 0.5), (0.5, 0.75), (0.75, 1)$, respectively. We chose this discretization since any non-empty interval in $\mathcal{E}^{(A)}$ or $\mathcal{E}^{(B)}$ can be represented as a union of these intervals (ignoring the interval endpoints).

We selected the Phase II total sample size, denoted $2n$, to be 528. This was, roughly, based on the sample size required for the fixed design (Figure 3.1a) to have type I error α and power $1 - \beta$, in the simple case of a constant conditional treatment effect $\Delta = \tau > 0$ and conditional variance σ^2. This sample size is $4\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2(\sigma^2/\tau^2)$; to maintain a realistic signal to noise ratio

in the Phase II trial, we set $\alpha = 0.1$, $\beta = 0.2$, $\sigma^2 = 9$ and $\tau = 0.554$ in this formula (and rounded down), which implies the total Phase II sample size is $2n = 528$. Our choices of α, β are typical for Phase II trials as described by Rubinstein et al. (2005). However, our choices of τ and σ were somewhat arbitrary; the value of τ was taken to be between the average treatment effect of 1 (under $\Delta = \delta_1$) and 0.25 (under each $\Delta = \delta_k, k = 2, 3, 4$). We set the prespecified sample size $N = 2473$ for each Phase III trial based on the above formula with parameters $\tau = 0.3$ (representing the minimum, clinically meaningful benefit), type I error $\alpha = 0.05$, type II error $\beta = 0.2$, and outcome $\sigma = 3$.

Our optimization problem is invariant to rescaling in that the optimal decision rules (as functions of the sufficient statistics $\tilde{S}^{(A)}, \tilde{S}^{(B)}$) are unchanged if we multiply all of n, σ^2, N by the same positive constant. For example, the optimal decision rules would be the same if we multiply these parameters by $1/4$, i.e., setting the Phase II total sample size $2n = 132$, outcome conditional variance $\sigma^2 = 9/4$, and Phase III sample size $N = 618$.

3.4.2 Optimal Adaptive Phase II Designs in Simulation Study

The optimal decision rules $d_{opt}^{(A)}$ and $d_{opt}^{(B)}$ are defined by (3.3), which depends on the prior π and the utility function U defined above. These rules are optimal for

a pair (π, U) in terms of the Bayes criterion (3.3). These rules may be suboptimal for any single state of nature δ_j, since the rules find the best tradeoff in expected utility U across these different states of nature, where the relative importance of each state of nature depends on π. The approximate optimal decision rules were computed using the backward induction method in Section 3.3.3.

We explored the performance of the optimal decision rules $(d_{opt}^{(A)}, d_{opt}^{(B)})$ by conducting simulated trials with data generated from one treatment effect function δ_k at a time. For each δ_k, we simulated 200 trials under $\Delta = \delta_k$ using the precomputed decision rules $(d_{opt}^{(A)}, d_{opt}^{(B)})$. The frequency of each population in $\mathcal{E}^{(A)}$ getting selected for Stage B enrollment was recorded; also recorded was the frequency of each population in $\mathcal{E}^{(B)}$ getting selected for the future Phase III trials. The optimal rules $d_{opt}^{(A)}$ and $d_{opt}^{(B)}$ depend on the prespecified π and U, and are determined prior to the start of the study, regardless of our setting Δ to be different curves $\delta_1, \ldots, \delta_k$, indicating the decisions to make according to the observed data during the course of the simulation study.

Table 3.1 shows the operating characteristics of the optimal, adaptive Phase II design $(d_{opt}^{(A)}, d_{opt}^{(B)})$ using the utility function (3.2) with $\lambda = 0.01, c = 0.32$. The top half of Table 3.1 shows the frequency of different choices for Stage B enrollment corresponding to $d_{opt}^{(A)}$. For example, 34% of the simulated trials recommend to enroll participants from the overall population in Stage B of the Phase II adaptive design when $\Delta = \delta_1$ is the data generating distribution. For δ_2 and δ_3, where only

half of the overall population benefits from the treatment, there is a 67% chance of enrolling at least the population who benefit from treatment in Stage B.

Table 3.1: Operating Characteristics of $d_{opt}^{(A)}$ (top) and $d_{opt}^{(B)}$ (bottom) for $\lambda = 0.01, c = 0.32$.

	Distribution of Population Recommended for Stage B Enrollment by $d_{opt}^{(A)}$			
	$d_{opt}^{(A)} = (0,1)$	$d_{opt}^{(A)} = (0.5,1)$	$d_{opt}^{(A)} = (0,0.5)$	$d_{opt}^{(A)} = (0.25,0.75)$
$\Delta = \delta_1$	0.34	0.27	0.24	0.15
$\Delta = \delta_2$	0.30	0.37	0.21	0.12
$\Delta = \delta_3$	0.34	0.22	0.33	0.11
$\Delta = \delta_4$	0.27	0.26	0.26	0.21
$\Delta = \delta_5$	0.67	0.14	0.13	0.06
$\Delta = \delta_6$	0.30	0.34	0.27	0.09

	Distribution of Population Recommended for Phase III trials by $d_{opt}^{(B)}$				
	$d_{opt}^{(B)} = (0,1)$	$d_{opt}^{(B)} = (0.5,1)$	$d_{opt}^{(B)} = (0,0.5)$	$d_{opt}^{(B)} = (0.25,0.75)$	$d_{opt}^{(B)} = \emptyset$
$\Delta = \delta_1$	**0.96**	0.01	0.01	0.02	0.00
$\Delta = \delta_2$	0.07	**0.53**	0.04	0.09	0.27
$\Delta = \delta_3$	0.11	0.02	**0.56**	0.09	0.22
$\Delta = \delta_4$	0.12	0.04	0.09	**0.40**	0.35
$\Delta = \delta_5$	0.00	0.00	0.00	0.00	**1.00**
$\Delta = \delta_6$	0.01	0.09	0.06	0.12	**0.72**

The bottom half of Table 3.1 shows the frequency of different choices for the population to enroll in the two, Phase III trials under the decision rule $d_{opt}^{(B)}$. We underline the number corresponding to the population who benefit from treatment (defined in Section 3.4.1) under the corresponding treatment effect function δ_k. We mark in bold the largest proportion in each row, which represents the population that is recommended most frequently for the future Phase III trials. For example, when data are generated under treatment effect function δ_2 (row 2), the proportion 0.53 is both underlined and in bold, which means that the corresponding popula-

tion $(0.5, 1)$ is the population who benefits and is the population most frequently recommended for Phase III trials. In every row in Table 3.1, the bold number coincides with the underlined number, which shows that in the plurality of simulated trials the optimal design chooses the population who benefits.

The results in Table 3.1 are for a particular choice of utility function parameters λ and c, which encode the relative importance of the cost of conducting Phase III trials and the treatment cost if the treatment is approved (including both financial costs and health costs such as side effects), respectively. We next show the impact of increasing c from 0.32 to 0.34, while holding all other parameters constant. Using

Table 3.2: Operating Characteristics of $d_{opt}^{(B)}$ for $\lambda = 0.01, c = 0.34$.

	Distribution of Population Recommended for Phase III trials by $d_{opt}^{(B)}$				
	$d_{opt}^{(B)} = (0, 1)$	$d_{opt}^{(B)} = (0.5, 1)$	$d_{opt}^{(B)} = (0, 0.5)$	$d_{opt}^{(B)} = (0.25, 0.75)$	$d_{opt}^{(B)} = \emptyset$
$\Delta = \delta_1$	**0.95**	0.01	0.02	0.02	0.00
$\Delta = \delta_2$	0.07	**0.48**	0.04	0.08	0.33
$\Delta = \delta_3$	0.09	0.03	**0.55**	0.08	0.25
$\Delta = \delta_4$	0.11	0.05	0.07	**0.39**	0.38
$\Delta = \delta_5$	0.00	0.00	0.00	0.00	**1.00**
$\Delta = \delta_6$	0.01	0.07	0.05	0.08	**0.79**

the utility function (3.2) with $\lambda = 0.01, c = 0.34$, we computed the operating characteristics of the component $d_{opt}^{(B)}$ of the optimal design, summarized in Table 3.2. There is a higher chance of making the decision not to conduct any Phase III trials (rightmost column), compared to Table 3.1, under each δ_k (except δ_5 where no Phase III trials are conducted with probability 1 in both tables). Under the global null hypothesis $\Delta = \delta_6$, the probability of conducting Phase III trials (which would

be a waste of resources) drops from 28% to 21% comparing $c = 0.32$ to $c = 0.34$. The tradeoff is that the optimal decision rule for $c = 0.34$ (Table 3.2) has lower probabilities of selecting the optimal population for enrollment in Phase III when the overall population or a subpopulation benefits ($\Delta \in \{\delta_1, \delta_2, \delta_3, \delta_4\}$).

3.4.3 Optimal Adaptive versus Fixed Phase II Trial Design

We solved the same optimization problem as in Section 3.4.2 using $\lambda = 0.01, c = 0.32$, except restricting to a fixed design, i.e., setting $d^{(A)}$ to be the constant function that always selects the full population $(0, 1)$ to enroll in Stage B of Phase II; only the function $d^{(B)}$ is optimized. The resulting design is referred to as the optimal fixed design. We compare its performance to that of the optimal adaptive Phase II design in order to determine the value added by adaptive enrichment, i.e., the value added by allowing enrollment to be restricted to a subset of the population in Stage B of Phase II.

The recommendation frequencies for Phase III trials for the optimal fixed design are very similar to those for the optimal adaptive design in Table 3.1, with the maximum difference between any 2 corresponding entries being 3%. Also, we computed the expected utility for the optimal adaptive design versus the fixed design, based on 15,000 simulated trials. The expected utility is 0.07 for both designs.

We also compared the contribution from each component of the utility function for these two designs, summarized in Table 3.3.

Table 3.3: Expected utility and expected value of its 4 components under the optimal fixed design and adaptive designs, respectively, based on the simulation setup in Section 3.4.1 using $\lambda = 0.01, c = 0.32$. Expectation is with respect to the prior π. The interval (r_l, r_u) in the formulas below represents the Phase III enrollment decision $d^{(B)}_{opt}(X)$.

	Fixed Design	Adaptive Design
Expected Utility $\mathbb{E}[U]$	0.07	0.07
Components of Expected Utility:		
Expected Treatment Benefit $\mathbb{E}[\int_{r_l}^{r_u} \Delta(r)dr]$	0.14	0.14
Expected Treatment Cost $\mathbb{E}[c(r_u - r_l)]$	0.10	0.10
P(2 Phase III Trials Conducted and Both Succeed)	0.27	0.27
Expected Phase III Cost $\mathbb{E}[\lambda 1(r_l \neq r_u)/(r_u - r_l)]$	0.82	0.80

The probability of having 2 successful Phase III trials is 0.27, averaged over the prior π. This is the probability of successfully demonstrating treatment efficacy when considering the combined Phase II/III trial sequence. This may seem like a low probability of the trial sequence succeeding. However, we next show that our choice of prior π implies that regardless of what is done during phases 2 and 3 (e.g., even if both sample sizes were arbitrarily increased), there is no way to achieve a probability greater than 41% of having a successful trial sequence. Intuitively, this is because the prior selects $\Delta = \delta_5$ or $\Delta = \delta_6$ with probability 0.6, and in such cases the treatment is not beneficial for any stratum of the baseline score.

We provide an upper bound for the probability of two successful Phase III trials under the prior π and an arbitrary rule for deciding which population to enroll in Phase III. Under π, for the event $\max_{r\in(0,1)} \Delta(r) \leq 0$, there is at most 0.05 probabil-

ity of each Phase III trial succeeding. Therefore, the probability of two successful Phase III trials (regardless of both the decision rule for Phase III enrollment and the Phase III sample size) is at most $1 - 0.6(1 - 0.05^2) = 0.4015$. By comparison, the corresponding 0.27 probability for the utility-maximizing designs above (where the objective function involves terms other than just power in Phase III) is not insubstantial.

3.4.4 Impact of Adaptive Design on Number Assigned to Superior Treatment During Phase II

In this section, we shift our focus to what happens to participants during Phase II, rather than after Phase II. Specifically, we measure the impact of being enrolled in the Phase II trial compared to not being enrolled; we assume not being enrolled in the trial means that a patient would have received the standard of care, i.e., the control. We focus only on Stage B participants in the Phase II trial since our goal is to contrast the fixed versus adaptive designs, and these designs have identical patient outcome distributions during Stage A.

We say that Stage B participant i with baseline score R_i in study arm T_i is assigned to a superior treatment (compared to control) if $\Delta(R_i) > 0$ and $T_i = 1$, that is, if the conditional treatment effect is positive in that participant's baseline stratum and she/he is assigned to the treatment arm $T_i = 1$. For example, if $\Delta = \delta_2$,

each participant i with baseline score $R_i > 0.5$ in arm $T_i = 1$ is assigned to a superior treatment. We denote the proportion of Stage B participants who are assigned to a superior treatment as $f_{\text{prop}} = \frac{1}{n}\sum_i 1\{T_i = 1, \Delta(R_i) > 0\}$, where the sum is over the n Stage B participants. Similarly, define the average benefit to Stage B participants of being enrolled in the trial compared to not being enrolled, as $f_{\text{ben}} = \frac{1}{n}\sum_i 1(T_i = 1)\Delta(\tilde{R}_i)$, where the sum is over the n Stage B participants.

Table 3.4 presents the expected proportion $\mathbb{E}(f_{\text{prop}}|\Delta)$ of Stage B participants who are assigned to a superior treatment and the expected average benefit $\mathbb{E}(f_{\text{ben}}|\Delta)$, respectively, conditional on the treatment effect function Δ. These expectations depend on the decision rule $d^{(A)}$. We evaluated the optimal rule $d^{(A)} = d^{(A)}_{opt}$ from Section 3.4.3, which was optimized for the utility function U in (3.2) using $\lambda = 0.01, c = 0.32$. We also evaluate the fixed (non-adaptive) decision rule $d^{(A)} \equiv (0, 1)$. The evaluation of these decision rules was based on the same set of simulations from Section 3.4.3.

For every treatment effect function δ_k where some but not all strata of r benefit from the treatment (i.e., for $\delta_k : k \in \{2, 3, 4\}$), the difference between the expected proportion assigned to a superior treatment $\mathbb{E}(f_{\text{prop}}|\Delta = \delta_2)$ for $d^{(A)}_{opt}$ versus the fixed design is 3-5%, as shown in Table 3.4. For the expected benefit f_{ben}, the relative improvement comparing adaptive versus fixed designs ranges from $(0.132 - 0.125)/0.125 \approx 5\%$ to $(0.151 - 0.125)/0.125 \approx 20\%$, as we consider different δ_k.

Table 3.4: Expected value of f_{prop} and f_{ben}, comparing fixed design versus adaptive design. Top half shows $\mathbb{E}(f_{\text{prop}}|\Delta = \delta_k)$ and bottom half shows $\mathbb{E}(f_{\text{ben}}|\Delta = \delta_k)$.

(a) $\mathbb{E}(f_{\text{prop}}|\Delta = \delta_k)$:

	Δ	Fixed Design Rule	Adaptive Design Rule
Data Generating Distributions	δ_1	50%	50%
	δ_2	25%	29%
	δ_3	25%	28%
	δ_4	25%	30%
	δ_5	0%	0%
	δ_6	0%	0%

(b) $\mathbb{E}(f_{\text{ben}}|\Delta = \delta_k)$:

	Δ	Fixed Design Rule	Adaptive Design Rule
Data Generating Distributions	δ_1	0.5	0.5
	δ_2	0.125	0.137
	δ_3	0.125	0.132
	δ_4	0.125	0.151
	δ_5	-0.5	-0.5
	δ_6	0.000	0.000

Despite not providing a higher expected utility than the optimal fixed design (as shown in Section 3.4.3), the adaptive design turned out to have advantages over the fixed design in the number of participants assigned to a superior treatment in the Phase II trial (which is not reflected in the utility function from Section 3.3.1). That is, although the original aim of the Phase II adaptive design was to provide more targeted information to assist in the decision at the end of Phase II, a by-product is that the optimal adaptive design can lead to more participants assigned to a superior treatment in Phase II. From this perspective, such adaptive designs may be more ethical to conduct (under the assumptions built into our simulation study).

3.5 Simulation Study Mimicking Features from the MISTIE II Trial

We applied our optimized Phase II designs in simulation studies where the data generating distribution mimics features from the completed MISTIE II trial. The baseline score R is the quantile of intracerebral hemorrhage (ICH) volume. We define $M = 4$ baseline score categories demarcated by the 25%, 50% and 75% percentiles of ICH volume, which correspond to 33ml, 43ml, 57ml, respectively. The discrete baseline scores $\tilde{R} = 1, 2, 3, 4$ correspond to the ranges of ICH volume (in ml) $[17, 33)$, $[33, 43)$, $[43, 57)$, $[57, 120]$, respectively. The data set consists of 90 participants with complete observations $(\tilde{R}_i, T_i, Y_i) : i = 1, \ldots, 90$ in the MISTIE trial. The primary outcome is the indicator of whether the participant's functional disability score, as measured by the modified Rankin Scale, is 3 or less.

We consider adaptive Phase II trial designs that are optimized for the problem setup in Section 3.4.1 using sample size $n = 64$ per stage, and utility function parameters $c = 0.1$ and each of $\lambda \in \{0.004, 0.01\}$; the resulting optimal designs are denoted $d_{opt}^{(A)}, d_{opt}^{(B)}$ (which differ for the two different values of λ used). These optimization problems did not incorporate any features of the MISTIE data except that the baseline score is discretized into 4 categories.

We investigated the performance of the above, optimized designs in simulated Phase II trials where the distribution of the outcome Y given study arm T and

baseline score $\tilde{R}$ is the empirical distribution in the MISTIE trial. The estimated treatment effects $\tilde{\Delta}(\tilde{r})$ from this distribution are -0.07 for $\tilde{r} = (0, 0.25)$, 0.22 for $\tilde{r} = (0.25, 0.5)$, 0.12 for $\tilde{r} = (0.5, 0.75)$ and 0.19 for $\tilde{r} = (0.75, 1)$.

Data in Stage A of each simulated Phase II trial were generated as follows: equal numbers from each baseline category were enrolled and assigned to each arm; for each participant i, a random outcome Y was drawn from the MISTIE data set empirical distribution conditioned on that participant's study arm and baseline score $(T_i, \tilde{R}_i)$. Stage B data were generated analogously, except enrollment was only from the selected population.

The above data generating distribution violates our assumption that the outcome is normally distributed conditional on the study arm and baseline score. Also, the induced treatment effect function is not in the support of the prior π. This provided an opportunity to evaluate the performance of our optimized Phase II designs in a scenario that differs from those it was optimized for.

Using the above data generating distribution, we simulated 200 Phase II trials and computed the proportion of trials with each population recommendation for Phase III. The results are summarized in Figures 3.4 and 3.5. In each figure, each rectangle's base represents the interval corresponding to a population recommendation in $\mathcal{E}^{(B)}$ for Phase III enrollment, and its height represents the observed proportion of that recommendation under the optimal design $d_{opt}^{(B)}$. In Figure 3.4, for example, the population $\tilde{R} \in (0.5, 1)$ (base) is recommended for Phase III en-

rollment in 30% (height) of simulated trials.

Figures 3.4 and 3.5, which differ only in the Phase III duration cost $\lambda = 0.004$ vs. $\lambda = 0.01$, show the impact of increasing this cost. This increase in λ incentivizes recommending a broader population for Phase III enrollment (to increase the enrollment rate) or recommending no Phase III trials. Comparing Figures 3.4 and 3.5, the overall population is recommended with frequencies 4.5% vs. 8%, and no Phase III trials are recommended with frequencies 17.5% and 32.5%, respectively.

Since we sampled data from the empirical distribution of the MISTIE II trial as described above, the true state of nature $\tilde{\Delta}(\tilde{r})$ is $(-0.07, 0.22, 0.12, 0.19)$ for baseline ICH quartiles $\tilde{r} = 1, 2, 3, 4$, respectively. The population who benefit from treatment, in terms of baseline ICH quartiles $\tilde{R}$, consists of the last 3 categories, which represent ICH volume quantiles $(0.25, 1]$. Our Phase II designs only allowed recommending a population to enroll in Phase III from the action set $\mathcal{E}^{(B)}$ defined in (3.6). This precluded enrolling the true population who benefit from treatment $(0.25, 1]$. The allowed enrollment choices were the following: $(0, 1)$, which has the disadvantage of including the first quartile who are slightly harmed by treatment; $(0.5, 1)$ or $(0.25, 0.75)$, each of which has the disadvantage of excluding a quartile who benefit; $(0, 0.5)$, which has both types of disadvantages; and the empty set. Each optimized decision rule $d_{opt}^{(B)}$ in Figures 3.4 and 3.5 selects one of the most desirable options, i.e., one of $(0.25, 0.75)$,$(0.5, 1)$,$(0, 1)$, with total probability ranging

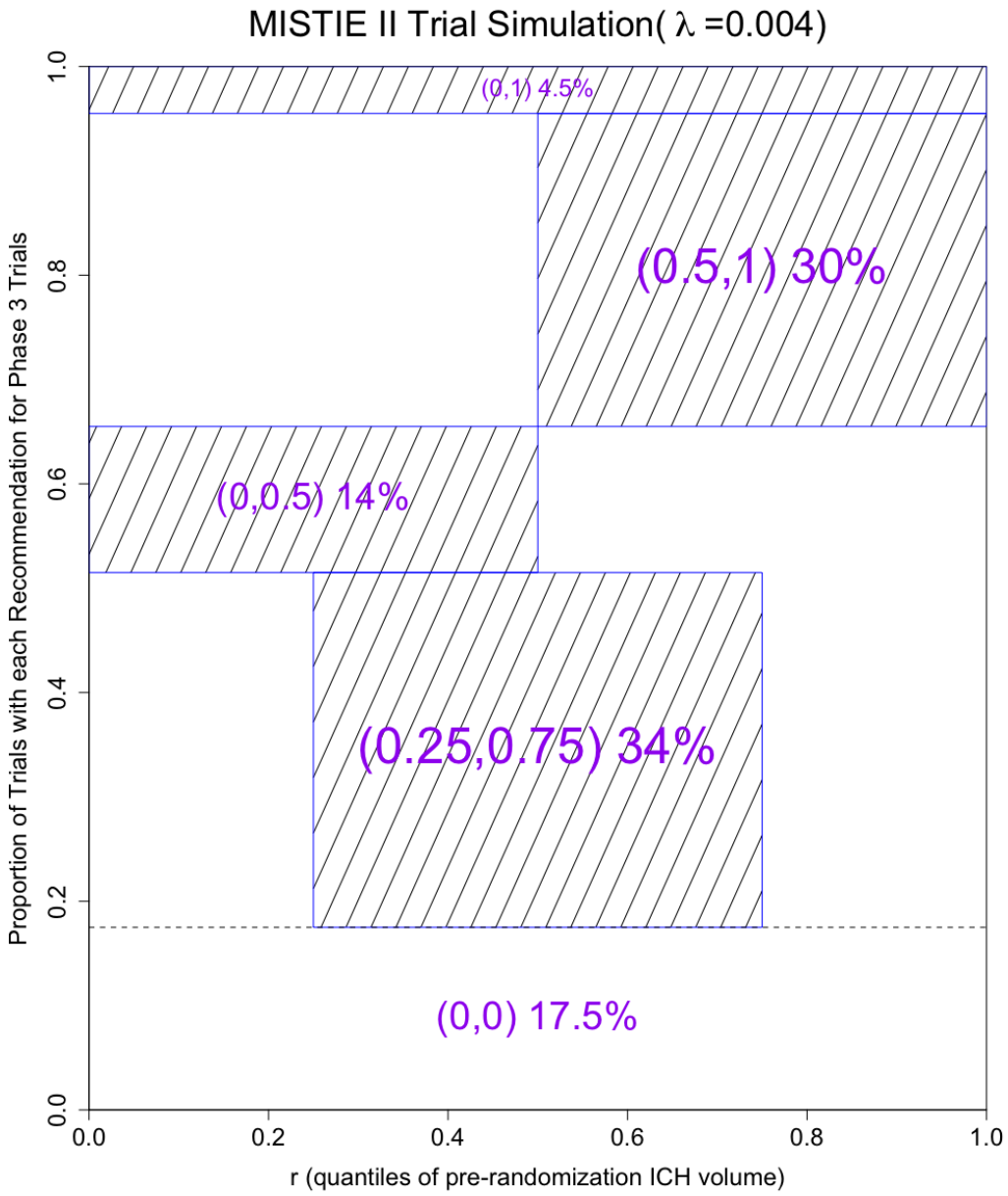


Figure 3.4: The plot shows the proportion of simulated trials in which the optimized adaptive Phase II design makes each recommendation in $\mathcal{E}^{(B)}$ for Phase III trial enrollment, at $\lambda = 0.004$.

from 60-68.5%.

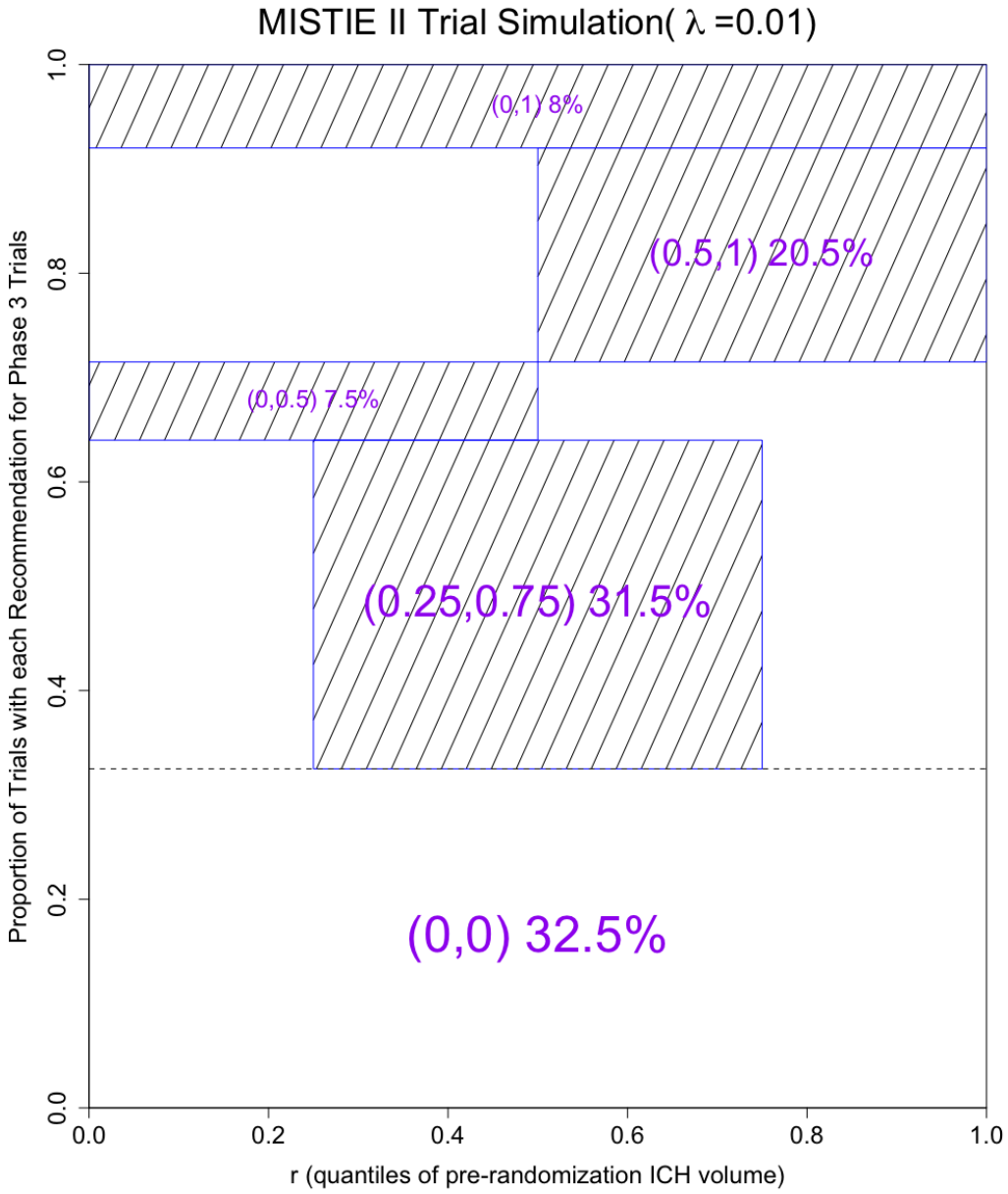


Figure 3.5: The plot shows the proportion of simulated trials in which the optimized adaptive Phase II design makes each recommendation in $\mathcal{E}^{(B)}$ for Phase III trial enrollment, at $\lambda = 0.01$.

3.6 Remarks

We optimized a Phase II adaptive enrichment design where the utility function U represents a combination of the cost of conducting Phase III trials (if they are recommended after Phase II) and the improved health of the future population due to treatment minus the treatment's cost (if the Phase III trials succeed). Despite not providing a higher expected utility than the optimal fixed design, the adaptive design turned out to have advantages over the fixed design in the number of participants assigned to a superior treatment in the Phase II trial. A limitation of our designs is that we require the primary outcome to be measured on each participant relatively soon after her/his enrollment, in order to avoid a long pause in enrollment between Stages A and B; this limitation is shared by many adaptive enrichment designs.

In Section 3.4.1, we discretized the continuous-valued baseline score R into the 4 level categorical variable $\tilde{R}$ by partitioning the support $(0, 1)$ of R into 4 equal length intervals. It may be possible to increase the expected utility of the optimal decision functions if we modify the problem by using a finer discretization of $(0, 1)$. Using a finer discretization, e.g., consecutive intervals of width $1/8$, could increase the information available about the unknown value of Δ; this could be used to improve decisions and increase expected utility. A tradeoff is that increasing the fineness of the discretization leads to increased variation and computational complexity. It is an area for future investigation to determine how much added value

a finer discretization provides, and at what computational cost.

The prior π used in the simulation studies is one possible choice; our general approach can be applied to an arbitrary set of weights and any finite set of conditional treatment effect functions, i.e., applied to any prior consisting of a finite set of point masses. The choice of prior would ideally be chosen based on prior scientific knowledge and data from earlier studies, aiming to provide an exhaustive list of possibilities. A possible future extension is to consider a mixture of such conditional treatment effect functions. The value of the parameters c and λ in the utility function (3.2) can be picked by considering the operating characteristics of the simulations conducted prior to the study as well as the subject expertise.

Another area for future research is to incorporate the possibility of early stopping for futility at the end of Stage A of Phase II, which could potentially save resources. We also could consider sample sizes in Stage B of Phase II that differ from the sample size in Stage A, or that are adaptively selected based on Stage A data. Similarly, we could consider setting the sample size for the Phase III trials based on the data from Phase II. Another area for future research is to consider a variety of different utility functions, e.g., incorporating the trial design cost function from Emerson et al. (2011).

The R code for our computations in Section 3.4 and 3.5 is available at `https://github.com/duyu8411/Phase2AdaptDesign`.

3.7 Appendix

3.7.1 Discretized versions of $\delta_1, \ldots, \delta_6$

Table 3.5 gives the discretized versions of $\delta_1, \ldots, \delta_6$, based on applying (3.1) to each.

Table 3.5: Average treatment effect $\tilde{\Delta}(\tilde{r})$ in each stratum $\tilde{r} \in \{1, 2, 3, 4\}$, under each possible $\Delta = \delta_1, \delta_2, \ldots, \delta_6$, as derived from Figure 3.3. In the row above the horizontal line, each stratum $\tilde{r}$ is followed by the interval of the baseline score that it represents.

	Average Treatment Effect $\tilde{\Delta}(\tilde{r})$ in Each Statum $\tilde{r}$ and Overall $(\mathbb{E}[\tilde{\Delta}(\tilde{R})])$				
	$\tilde{r} = 1; (0, 0.25)$	$\tilde{r} = 2; (0.25, 0.5)$	$\tilde{r} = 3; (0.5, 0.75)$	$\tilde{r} = 4; (0.75, 1)$	Overall
$\Delta = \delta_1$	1	1	1	1	1
$\Delta = \delta_2$	0	0	0.25	0.75	0.25
$\Delta = \delta_3$	0.75	0.25	0	0	0.25
$\Delta = \delta_4$	0	0.5	0.5	0	0.25
$\Delta = \delta_5$	-1	-1	-1	-1	-1
$\Delta = \delta_6$	0	0	0	0	0

3.7.2 Proof of (3.5)

We prove (3.5) from Section 3.3.3.

$$P(\Delta = \delta_k \mid X) \propto P(X \mid \Delta = \delta_k)P(\Delta = \delta_k) \tag{3.7}$$

$$\propto P(\widehat{\Delta}^{(B)}, d^{(A)}(X^{(A)}), \widehat{\Delta}^{(A)} \mid \Delta = \delta_k)P(\Delta = \delta_k) \tag{3.8}$$

$$= P^E(\widehat{\Delta}^{(B)}, \widehat{\Delta}^{(A)} \mid \Delta = \delta_k)P(\Delta = \delta_k) \tag{3.9}$$

$$= P^E(\widehat{\Delta}^{(B)} \mid \Delta = \delta_k)P^E(\widehat{\Delta}^{(A)} \mid \widehat{\Delta}^{(B)}, \Delta = \delta_k)P(\Delta = \delta_k) \tag{3.10}$$

$$= P^E(\widehat{\Delta}^{(B)} \mid \Delta = \delta_k)P^E(\widehat{\Delta}^{(A)} \mid \widehat{\Delta}^{(B)})P(\Delta = \delta_k) \tag{3.11}$$

$$\propto P^E(\widehat{\Delta}^{(B)} \mid \Delta = \delta_k)P(\Delta = \delta_k), \tag{3.12}$$

where (3.7) follows from Bayes' rule; (3.8) holds since $\tilde{S}^{(B)}$ is a sufficient statistic for Δ; (3.9) follows from the sequential structure of the data generating process; (3.10) follows from the definition of conditional probability; (3.11) follows since $\widehat{\Delta}^{(B)}$ is a sufficient statistic for Δ in the design that always enrolls population E in Stage B; (3.12) holds because $P^E(\widehat{\Delta}^{(A)} \mid \widehat{\Delta}^{(B)})$ does not depend on Δ, and thus is absorbed into the proportionality constant that depends only on X.

3.7.3 Computation of $d_{opt}^{(A)}$

At the end of Stage A, we have collected Stage A data $X^{(A)}$. The goal at this point in the backward inductive computation is to maximize the expected utility

conditioned on $X^{(A)}$ over all possible enrollment choices at the end of Stage A, $\mathcal{E}^{(A)}$, assuming that $d_{opt}^{(B)}$ will subsequently be used after Stage B; the maximizer is defined as $d_{opt}^{(A)}(X^{(A)})$. It follows from (3.3) that

$$\begin{aligned} d_{opt}^{(A)}(X^{(A)}) &= \underset{d^{(A)}}{\operatorname{argmax}}\, \mathbb{E}\left[U\left\{ d_{opt}^{(B)}(X); \Delta \right\} \middle| X^{(A)} \right] \\ &= \underset{d^{(A)}}{\operatorname{argmax}}\, \mathbb{E}\left[U\left\{ d_{opt}^{(B)}\left(X^{(A)}, X^{(B)}[d^{(A)}\{X^{(A)}\}] \right); \Delta \right\} \middle| X^{(A)} \right], \end{aligned} \tag{3.13}$$

where the expectation is with respect to the conditional distribution of $(\Delta, X^{(B)})$ given the Stage A data $X^{(A)}$ (which is induced by the prior π). Analogous to the computation of $d_{opt}^{(B)}$, we use Monte-Carlo simulation to approximate the posterior conditional expectation in (3.13) via the posterior distribution of $(\Delta, X^{(B)})$ given $X^{(A)}$, where we first draw Δ from the posterior distribution $P\left(\Delta | X^{(A)}\right)$ via (3.12), and generate the Stage B data, $X^{(B)}$, given Δ and an enrollment choice $d^{(A)}(X^{(A)}) \in \mathcal{E}^{(A)}$, according to the data generating distribution specified in Section 3.2.2. We thus obtain the full data, $X = (X^{(A)}, X^{(B)})$, and a value for the utility function. The average values of the utility function from multiple such draws of data approximates the posterior conditional expectation for this particular enrollment choice. We consider each possible enrollment choice in $\mathcal{E}^{(A)}$, and set the optimal enrollment decision $d_{opt}^{(A)}(X^{(A)})$ to be the maximizer over $\mathcal{E}^{(A)}$ of this posterior conditional expectation.

Chapter 4

Phase III Studies: Bias, Variance, and Sample Size Reductions due to Adjustment in Adaptive Enrichment Designs

SUMMARY.[1]

In adaptive enrichment designs, early stopping of a subpopulation with sufficient evidence of treatment efficacy, futility or harm is allowed according to preplanned rules for modifying enrollment criteria, while the remaining subpopula-

[1]This Chapter 4 is adapted from the working paper "**Yu Du**, Tianchen Qian, Huitong Qiu, Michael Rosenblum. *Bias, Variance, and Sample Size Reductions due to Adjustment for Prognostic Baseline Variables and Short Term Outcomes in Adaptive Enrichment Trial Designs with Delayed Outcomes.*"

tions continue to be enrolled. Most existing methods for constructing adaptive enrichment designs are limited to situations where patient outcomes are observed soon after enrollment. This is a major barrier to the use of such designs in practice, since for many diseases the outcome of most clinical importance does not occur shortly after enrollment. For the Phase III studies, we propose to use semiparametric, locally efficient estimators, at each analysis of adaptive enrichment designs for delayed outcome, leveraging information in baseline variables and short-term outcomes to improve precision. We evaluate power, expected sample size, bias, variance, and mean squared error for our design and compare with a non-adaptive design and unadjusted estimator, through simulations of a trial for a new surgical intervention for stroke. We strongly control the familywise Type I error rate, asymptotically.

4.1 Background

We address the problem of designing a confirmatory randomized trial of an experimental treatment versus control when the primary outcome is measured with delay and there are multiple subpopulations of interest. Our methods were developed to solve a problem in designing a trial for a new surgical treatment for stroke, where there are two subpopulations of interest and outcomes are measured a fixed time (180 days) from enrollment. However, our general method applies to larger

numbers of subpopulations and to time-to-event outcomes.

We propose to use semiparametric, locally efficient estimators, at each analysis of adaptive enrichment designs for delayed outcome, leveraging information in baseline variables and short-term outcomes to improve precision. To illustrate our approach, consider an interim analysis that occurs just after 50% of a trial's total enrollment. Due to delayed outcomes, less than 50% of final (i.e., primary) outcomes will be observed. However, all enrolled participants will have baseline variables observed, some will have short-term outcomes observed, and a further subset will have the final outcome observed. If the short-term outcomes and baseline variables are correlated with the final outcome, they can provide valuable information that we harness through the estimators described below. For example, in the stroke trial that motivated this work (described in Section 4.2), baseline stroke severity and 30-day disability score are strongly correlated with the primary outcome of disability score at 180 days.

In our context of a randomized trial, these semiparametric, locally efficient estimators converge to the true average treatment effect, without having to make any parametric model assumptions. As stated in the Patient-Centered Outcomes Research Institute Methodology Report (PCORI, 2013, Section 9), the chief statistical concerns for adaptive designs include "type I error, power, and sample size distributions, as well as the precision and bias in the estimation of treatment effects". Similar concerns are emphasized in the FDA draft guidance on adaptive designs

for drugs and biologics (FDA, 2010) and for medical devices (FDA, 2015). We therefore evaluate power, expected sample size, maximum sample size, bias, variance and mean squared error for the proposed estimator and adaptive enrichment design, comparing to standard, unadjusted estimator and non-adaptive design in simulations designed to mimic key features of a completed stroke trial. The unadjusted estimator ignores short-term outcomes and baseline variables. The simulation witnesses tangible improvements in precision, a 19-20% reduction in both expected sample size and maximum sample size, and comparable power, bias, variance and mean squared error, over unadjusted estimator.

Our designs strongly control the familywise Type I error rate as required, e.g., by the U.S. Food and Drug Administration in their draft guidance on adaptive designs for drugs and biologics (FDA, 2010). This means that the probability of rejecting one or more true null hypotheses is at most the desired level, under any data generating distribution.

A general method for ensuring strong control of the familywise Type I error rate is the p-value combination approach of Bauer (1989), Bauer and Köhne (1994), Lehmacher and Wassmer (1999), which has been applied to adaptive enrichment designs by, e.g., Bretz et al. (2006); Schmidli et al. (2006); Jennison and Turnbull (2007); Brannath et al. (2009). Another general method for achieving this goal is the conditional error function approach of Müller and Schäfer (2001), which has been applied to adaptive enrichment designs by Friede et al. (2012). Both of the

above methods require assumptions that do not generally hold when using semiparametric, locally efficient estimators in our context, as described in Section 4.4. To take advantage of precision gains that these estimators make possible, we use a multiple testing procedure proposed by Rosenblum et al. (2016) for adaptive enrichment designs that takes full advantage of correlations among related statistics, including statistics for the same population at different times, and statistics for different but overlapping populations.

If information for the subpopulations of interest is low throughout the trial (e.g., if accrual is slower than projected), the trial reverts to a standard, group sequential trial for the overall population. This feature protects against the realistic possibility of insufficient information at interim analyses to make an informed decision regarding changing enrollment criteria.

We describe our motivating application in Section 4.2. The general problem is defined in Section 4.3. In Section 4.4, we present semiparametric, locally efficient estimators that are used in our designs. The general framework of adaptive enrichment designs for delayed outcomes is introduced in Section 4.5. We apply our designs to the stroke trial application in Section 4.6. An evaluation of bias, variance and mean squared error for the proposed estimator and adaptive enrichment design is given in Section 4.7. Section 4.8 gives limitations of our approach and areas for future research.

4.2 Motivating Application: MISTIE stroke trial

As introduced in Chapter 3, Minimally-Invasive Surgery Plus rt-PA for Intracerebral Hemorrhage Evacuation, abbreviated as MISTIE (Morgan et al., 2008), is a new surgical treatment for intracerebral hemorrhage (ICH). Unlike the Phase II studies in 3, we consider the problem of planning a Phase III trial to compare the MISTIE surgical treatment to the standard of care, assessing if it provides more treatment benefit. A binary disability score, as measured by the modified Rankin Scale (mRS) at 180 days from enrollment, serves as the primary outcome where a mRS score of 3 or less is considered a success. The difference between the probability of success comparing MISTIE treatment to standard of care (control) is therefore defined to be the average treatment effect.

Prior data indicated greater uncertainty of the treatment effect for the subpopulation of participants with large (at least 10ml) intraventricular hemorrhage (IVH) at baseline, called large IVH participants. All others are called small IVH participants. The clinical investigators conjectured that two scenarios were most likely to occur if the treatment is beneficial:

1. Treatment benefits both subpopulations.

2. Treatment benefits only subpopulation with small IVH.

We consider the adaptive enrichment designs for testing the corresponding two null hypotheses: the MISTIE treatment provides no benefit for the overall population as well as for the small IVH subpopulation.

4.3 General Problem Definition

4.3.1 Subpopulations and Data Structure for Each Participant

We assume the overall population is partitioned into m disjoint subpopulations, which are functions of variables measured before randomization. For each $s \in \{1, \ldots, m\}$, let p_s denote the proportion of the overall population in subpopulation s, which we assume is known. In all our designs, subpopulation definitions must be specified before the trial starts. We do not consider the more challenging problem of selecting from a set of candidate biomarkers to define subpopulations of interest using data accrued in an ongoing trial as in, e.g., (Freidlin and Simon, 2005).

Each participant i, when followed up completely, has full data vector

$$D_i = (S_i, W_i, A_i, L_i^{(1)}, \ldots, L_i^{(T)}, Y_i),$$

where S_i denotes subpopulation, W_i is a vector of baseline (pre-randomization) variables, A_i is the treatment indicator ($A_i = 1$ indicates treatment and $A_i = 0$ indicates control), $L_i^{(1)}, \ldots, L_i^{(T)}$ are variables observed after randomization, and Y_i is the final (i.e., primary) outcome. We assume that $L_i^{(1)}, \ldots, L_i^{(T)}, Y_i$ are observed at preplanned durations $d_1, \ldots, d_T, d_Y$, respectively, from the time of enrollment, such that $0 < d_1 < \cdots < d_T < d_Y$. The subscript i is omitted when referring to a generic participant.

A special case of interest is where $L^{(1)}, \ldots, L^{(T)}$ represent the the same quantity as in the primary outcome, but measured at the earlier times $d_1, \ldots, d_T$; we refer to $L^{(1)}, \ldots, L^{(T)}$ as short-term outcomes, though in general they can be any variables measured after randomization. For example, in the MISTIE trial we have $T = 1$ and the following data are measured for each participant: subpopulation $S \in \{1, 2\}$ (small IVH and large IVH participants, respectively); baseline variables $W =$ (NIH Stroke Scale, clot volume, and Glasgow Coma Scale); treatment indicator A; indicator $L^{(1)}$ of functional disability score (mRS) ≤ 3 at 30 days from enrollment; the primary outcome Y, which is the indicator of mRS ≤ 3 at 180 days from enrollment.

4.3.2 Interim Analyses

The timing of interim analyses can be any preplanned function of calendar time and/or information accrued. Also preplanned are the maximum number of stages

K and the maximum, cumulative sample size $N_{s,\max}$ for each subpopulation s. At the start of the trial, all subpopulations are continuously enrolled. Enrollment of subpopulation s continues until either $N_{s,\max}$ have been enrolled or the preplanned rule for enrollment restriction causes enrollment to cease. The latter case is called an early stop for subpopulation s (as opposed to stopping enrollment once the maximum sample size $N_{s,\max}$ is reached). Enrollment from a subpopulation cannot be restarted if it has stopped. We assume the first interim analysis time is set so that at least $n_{\min}$ participants from each subpopulation have final outcomes observed.

We assume subpopulation enrollment rates are proportional to subpopulation sizes, i.e., if enrollment has not stopped for subpopulations s and s', the ratio of the cumulative number enrolled from subpopulation s and s' is $p_s/p_{s'}$.

4.3.3 Assumptions on Data Generating Distribution

The subpopulations enrolled in each stage depend on data from previous stages and the decision rule for modifying enrollment criteria. For each participant i from subpopulation s, we assume his/her baseline data W_i is a random draw from an unknown distribution $Q_s^{(W)}$, independent of the data from all previously enrolled participants. We assume each participant is randomized with probability $1/2$ to each study arm, independent of subpopulation S and baseline variables W, i.e., $P(A = 1|S, W) = 1/2$; we call this the randomization assumption. We assume that for each participant i enrolled during stage k, conditioned on his/her subpopula-

tion S_i and treatment assignment A_i, we have $(W_i, L_i^{(1)}, \ldots, L_i^{(T)}, Y_i)$ is a random draw from an unknown distribution $Q_{S_i A_i}$ that is independent of the data from all other participants enrolled at or before stage k.

The sets of unknown distributions are denoted by $Q = \{Q_{sa} : s = 1, \ldots, m; a = 0, 1\}$ and $Q^{(W)} = \{Q_s^{(W)} : s = 1, \ldots, m\}$. We make no parametric model assumptions about these distributions. Nor do we assume that $L^{(1)}, \ldots, L^{(T)}$ are surrogates for the primary outcome Y. We assume a nonparametric model where the only assumptions are that $P(A = 1|S, W) = 1/2$ by randomization, and that Q and $Q^{(W)}$ satisfy regularity conditions such as being dominated by σ-finite measures ν and $\nu^{(W)}$, respectively, as described in Section 4.9.1 of the Appendix.

4.3.4 Definitions of Treatment Effects and Hypotheses

For each $s \in \{1, \ldots, m\}$, define the average treatment effect for subpopulation s as

$$\delta_s = E(Y|A = 1, S = s) - E(Y|A = 0, S = s).$$

For each $j \in \{0, \ldots, J\}$, define $\tilde{\mathcal{S}}_j \subseteq \{1, \ldots, m\}$ to represent the jth composite population of interest, consisting of the union of subpopulations in $\tilde{\mathcal{S}}_j$. The overall population, which includes all subpopulations, will generally be a composite population of interest, and we denote it by $\tilde{\mathcal{S}}_0 = \{1, \ldots, m\}$. The average treatment

effect in composite population $\tilde{\mathcal{S}}_j$ is defined as

$$\Delta_j = E(Y|A=1, S \in \tilde{\mathcal{S}}_j) - E(Y|A=0, S \in \tilde{\mathcal{S}}_j) = \sum_{s \in \tilde{\mathcal{S}}_j} p_s \delta_s \Big/ \sum_{s \in \tilde{\mathcal{S}}_j} p_s.$$

Throughout, δ represents the treatment effect in a subpopulation, and Δ represents the treatment effect in a composite population. For each $j \in \{0, \ldots, J\}$, let H_{0j} denote the null hypothesis of no average benefit of treatment compared to control for composite population $\tilde{\mathcal{S}}_j$, that is $H_{0j} : \Delta_j \leq 0$; the corresponding alternative hypothesis is $\Delta_j > 0$. We are interested in testing the set of null hypotheses $\{H_{0j} : j = 0, \ldots, J\}$.

4.3.5 Censoring

For each participant i and stage k, we define censoring variables that indicate which subset of the full data $(S_i, W_i, A_i, L_i^{(1)}, \ldots, L_i^{(T)}, Y_i)$ is observed by the end of stage k. Let $C_{i,k}^{(0)}$ denote the indicator of being enrolled during or prior to stage k; if $C_{i,k}^{(0)} = 1$, then at least (S_i, W_i, A_i), which are recorded when participant i is enrolled, are observed by the end of stage k. Let $C_{i,k}^{(t)}$ be the indicator that $L_i^{(t)}$ is observed by the end of stage k, and let $C_{i,k}^{(T+1)}$ be the indicator of observing Y_i by the end of stage k. In the special case of the MISTIE trial, for any participant i and stage k, the vector $(C_{i,k}^{(0)}, C_{i,k}^{(1)}, C_{i,k}^{(2)})$ has one of the following forms:

- $(0, 0, 0)$: no data observed, i.e., not yet enrolled by end of stage k;

- $(1, 0, 0)$: subpopulation S, baseline variables W and treatment indicator A observed;
- $(1, 1, 0)$: S, W, A, and short-term outcome $L^{(1)}$ observed;
- $(1, 1, 1)$: complete data vector $(S, W, A, L^{(1)}, Y)$ observed.

In general, we assume a monotone missingness structure, i.e., that $C_{i,k}^{(t)} \geq C_{i,k}^{(t+1)}$ for each $t \in \{0, \ldots, T\}, k \leq K, i \leq n$, and that $C_{i,k}^{(t)} \leq C_{i,k+1}^{(t)}$ for each $t \in \{0, \ldots, T+1\}, k \leq K-1, i \leq n$.

For clarity of presentation, throughout we assume the only cause of missing data is administrative censoring due to some participants not yet having experienced short-term and/or final outcomes at an interim analysis. In Section 4.8 we describe how to incorporate additional right censoring due to loss to follow-up, under the assumptions of censoring at random and that the censoring distribution can be correctly modeled.

4.3.6 Unadjusted Estimator

For a given population, the unadjusted estimator of the average treatment effect is defined to be the difference between sample means of the final outcome Y comparing those assigned to $A = 1$ versus $A = 0$. The unadjusted estimator at the end of stage k uses only data from participants who have Y observed at or before that time. For each subpopulation $s \in \{1, \ldots, m\}$, the unadjusted estimator of δ_s at

the end of stage k is

$$\hat{\delta}_{s,k,unadj} = \frac{\sum_i Y_i C_{i,k}^{(T+1)} 1\left[A_i = 1, S_i = s\right]}{\sum_i C_{i,k}^{(T+1)} 1\left[A_i = 1, S_i = s\right]} - \frac{\sum_i Y_i C_{i,k}^{(T+1)} 1\left[A_i = 0, S_i = s\right]}{\sum_i C_{i,k}^{(T+1)} 1\left[A_i = 0, S_i = s\right]},$$

where $1[X]$ is the indicator variable taking value 1 if X is true and 0 otherwise. The unadjusted estimator of treatment effect Δ_j for composite population $\tilde{\mathcal{S}}_j$ at stage k is $\hat{\Delta}_{j,k,unadj} = \sum_{s \in \tilde{\mathcal{S}}_j} p_s \hat{\delta}_{s,k,unadj} \Big/ \sum_{s \in \tilde{\mathcal{S}}_j} p_s$.

4.4 Semiparametric, Locally Efficient Estimators that Adjust for Baseline Variables and Short-term Outcomes

To take advantage of prognostic information in baseline variables and short-term outcomes, we use estimators that build on the general theory of semiparametric, local efficiency of Robins and Rotnitzky (1992). The main advantage of these estimators versus the unadjusted estimator is that when baseline variables and short-term outcomes are strongly correlated with the final outcome, as in the MISTIE trial example, these estimators can have greater precision than standard estimators. This allows for more informed decisions at interim analyses, improved power, or reduced expected sample size, with comparable bias, variance and mean

squared error. We use a targeted maximum likelihood estimator (TMLE) for longitudinal data developed by van der Laan and Gruber (2012) and implemented in the R package **ltmle** (Schwab et al., 2014); this estimator combines features of the general targeted maximum likelihood template of van der Laan and Rubin (2006) with the sequential regression approach of (Robins, 2000; Bang and Robins, 2005). We call this the adjusted estimator. Let $\hat{\Delta}_{j,k,adj}$ denote the adjusted estimator of Δ_j based on all data from participants in population $\tilde{\mathcal{S}}_j$ collected up to and including stage k. The precise definition of this estimator is given in Section 4.9.2 of the Appendix. It is also possible to use the semiparametric, locally efficient estimators of, e.g., Lu and Tsiatis (2011); Rotnitzky et al. (2012); Gruber and van der Laan (2012); Parast et al. (2014); Zhang (2015), as we discuss in Section 4.8.

To provide intuition for the adjusted estimator, consider the case of $T = 1$, so that each participant's data is a vector $(S, W, A, L^{(1)}, Y)$. The adjusted estimator accounts for chance imbalances in short-term outcomes between the following two groups: those with observed final outcomes, and the larger group of those with at least short-term outcomes measured; for example, if the short-term outcomes in this larger group predict worse final outcomes are coming in the pipeline, the estimator will adjust to reflect this. This adjustment is based on the working model fits that use data on participants with both short-term and long-term outcomes to estimate how predictive the former is for the latter. A similar adjustment is done for chance imbalances in baseline variables between those with at least short-

term outcomes measured and those with only baseline variables and study arm assignment measured. The above idea and estimator we use are not new (though the software implementing this TMLE estimator only recently became available). What is new is our proposing for a general framework to apply this in adaptive enrichment designs, and we show this leads to tangible improvements in precision through simulations in Section 4.6. We also show the impact of such estimators on power, expected sample size, maximum sample size, bias, variance and mean squared error.

The adjusted estimator involves working models, e.g., models for the mean of Y given $(L^{(1)}, A, W, S \in \tilde{\mathcal{S}}_j)$ for each $j \leq J$, which are fit using data accrued at a given interim analysis. These are called working models since we do not assume the true, unknown data generating distribution Q satisfies any of the assumptions of these models. For example, we use logistic regression models as working models, but do not assume the conditional distribution of Y given $(L^{(1)}, A, W, S)$ has the functional form of a logistic regression model. Under regularity conditions given in Section 4.9.1 of the Appendix, the adjusted estimator is consistent; this holds regardless of whether the working models are correctly specified. If the working models are correctly specified, then the adjusted estimator achieves the semiparameteric efficiency bound; this is the local efficiency property of the estimator.

When our discussion applies to a generic estimator, we suppress adj and $unadj$ in the subscript. For a given estimator $\hat{\Delta}_{j,k}$ of Δ_j, define the corresponding Wald

statistic $Z_{j,k} = \hat{\Delta}_{j,k}/\text{Var}(\hat{\Delta}_{j,k})^{1/2}$, where $\text{Var}(\hat{\Delta}_{j,k})$ is the variance of $\hat{\Delta}_{j,k}$. For a given population $\tilde{\mathcal{S}}_j$, the statistics $Z_{j,1}, \ldots, Z_{j,K}$ for the adjusted estimator do not generally have the canonical covariance matrix that arises when estimators (rescaled by the corresponding information) have the independent increments property described in (Scharfstein et al., 1997; Jennison and Turnbull, 1999). This means that the combination test approach and conditional error function approach (see references in Section 4.1) cannot be directly applied, since they assume this property or more generally the so-called p-clud property, neither of which holds for the adjusted estimator. This is shown in Section 4.9.1 of the Appendix.

4.5 Adaptive Enrichment Designs

We use the general framework proposed by Rosenblum et al. (2016) to define a new adaptive enrichment design, where we employ the error spending function approach to construct a set of group sequential boundaries for testing the corresponding null hypothesis (Rosenblum et al., 2016, Section 3.2), extending the approach of Slud and Wei (1982); Lan and DeMets (1983) to multiple populations.

Consider an estimator $\hat{\Delta}_{j,k}$ and its corresponding Wald statistic $Z_{j,k}$. We assume $\mathbf{Z} = \{Z_{j,k}\}_{j \leq J, k \leq K}$ is a multivariate normal family with covariance matrix Σ, satisfying for all $k, EZ_{j,k} \leq 0$ whenever H_{0j} is true. This assumption holds asymptotically if $\{\hat{\Delta}_{j,k}\}_{j \leq J, k \leq K}$ are consistent, asymptotically normal estimators,

such as the unadjusted and adjusted estimators described in Section 4.4. Define $\mathbf{Z}' = \{Z'_{j,k}\}_{j \leq J, k \leq K}$ to be a multivariate normal family of random variables with covariance Σ and all components having zero mean. Then $\{Z_{j,k} - EZ_{j,k}\}_{j \leq J, k \leq K}$ has the same joint distribution as $\mathbf{Z}'$. For clarity, in this subsection we assume Σ is known.

Let α^* denote the desired upper bound on the familywise Type I error rate, e.g., $\alpha^* = 0.025$ (since we use one-sided tests). Define error spending functions $\alpha_j(t)$ for each $j \in \{0, \ldots, J\}$ that are nondecreasing functions from the nonnegative reals to $[0, \alpha^*]$ that take the value 0 at $t = 0$ and satisfy $\sum_{j=0}^{J} \alpha_j(t) \leq \alpha^*$ for all $t > 0$. For each null hypothesis H_{0j}, at each interim analysis k, let $\mathcal{I}_{j,k} = 1/\text{Var}(\hat{\Delta}_{j,k})$ denote the information corresponding to the estimator $\hat{\Delta}_{j,k}$. Define the increment $\alpha_{j,k} = \alpha_j(\mathcal{I}_{j,k}/\mathcal{I}_{j,\max}) - \alpha_j(\mathcal{I}_{j,k-1}/\mathcal{I}_{j,\max})$, where $\mathcal{I}_{j,\max}$ is a predefined maximum information level for population $\tilde{\mathcal{S}}_j$, and $I_{j,0} = 0$ for all j. We define information $\mathcal{I}_{j,k}$ with respect to a particular estimator since this is the relevant quantity for decisions about interim monitoring when using such an estimator.

Let R_k denote the set of null hypotheses that have been rejected by the end of interim analysis k, and let $R_0 = \emptyset$. These sets are nested, i.e., $R_0 \subseteq R_1 \subseteq \cdots \subseteq R_K$. The set R_K represents all null hypotheses that are rejected by the end of the trial.

We let the rule r_k be any prespecified, measurable function from the data accrued up to and including stage k, to the set of subpopulations to enroll during stage $k + 1$. An example is given in Section 4.6.3. We assume that once enrollment

for a population has been stopped, it cannot be restarted. Let E_k denote the subset of null hypotheses $H_{01}, \ldots, H_{0J}$ for which all component subpopulations were enrolled through the end of stage k, i.e.,

$$E_k = \{H_{0j} : \text{for each } s \in \tilde{\mathcal{S}}_j, \text{ subpopulation } s \text{ is enrolled through the end of stage } k\}.$$

Multiple Testing Procedure M, as stated in Rosenblum et al. (2016):

At each interim analysis $k \leq K$, for each j in $0, \ldots, J$ in turn:

1. Define the threshold $u_{j,k}$ to be the solution to the equation such that:

$$P_\Sigma \left(\text{for all } k' \leq k, j' < j : Z'_{j',k'} \leq u_{j',k'}; \text{ and } Z'_{j,k} > u_{j,k}\right) = \alpha_{j,k}. \tag{4.1}$$

2. Reject H_{0j} if $H_{0j} \notin R_{k-1}$, $j \in E_k$, and $Z_{j,k} > u_{j,k}$.

The left side of (4.1) can be computed using the multivariate normal distribution function, e.g., implemented in the mvtnorm R package of Genz et al. (2014), which takes as input Σ. Given the previously computed values $\{u_{j',k'}\}_{k' \leq k, j' < j}$, the solution $u_{j,k}$ to (4.1) can be computed to high precision by the bisection (binary search) method. Rosenblum et al. (2016) has proved that for any early stopping rule (adaptive enrichment rule) r_k (even modified for any reason during an ongoing trial), multiple testing procedure M strongly controls the familywise Type I error rate at level α^*.

Above we considered the case where Σ, is known. In practice, Σ will be estimated as the trial progresses, just as in the error spending approach for a single

null hypothesis in a standard, group sequential design. The covariance matrix Σ can be consistently estimated with the nonparametric bootstrap. We describe a procedure for this in Section 4.9.3 of the Appendix.

4.6 Simulations

4.6.1 Overview

Consider the problem of planning the Phase III MISTIE trial, as introduced in Section 4.2. The variables $(S, W, A, L^{(1)}, Y)$ defined in the third paragraph of Section 4.3 are measured for each participant. We refer to those with small IVH as subpopulation 1, and those with large IVH as subpopulation 2. The composite populations of interest are the combined population, denoted by $\tilde{\mathcal{S}}_0 = \{1, 2\}$, and subpopulation 1, denoted by $\tilde{\mathcal{S}}_1 = \{1\}$. We test the corresponding null hypotheses H_{00} and H_{01}. In the adaptive enrichment design literature, it is not uncommon to focus on the null hypotheses for a single subpopulation and the combined population, e.g., Wang et al. (2007); Brannath et al. (2009); Jenkins et al. (2011); Boessen et al. (2013), Stallard et al. (2014, Section 5). We assume $p_1 = 1/3$ based on prior studies (Hanley, 2012). We assume the enrollment rate to be 50 patients per year for subpop 1, and 100 for subpop 2, based on projected enrollment rates for the MISTIE Phase III trial.

The clinical investigators in the MISTIE trial were interested in the following three scenarios: (a) $\delta_1 = 12.2\%$, $\delta_2 = 12.2\%$; (b) $\delta_1 = 12.2\%$, $\delta_2 = 0\%$; (c) $\delta_1 = \delta_2 = 0$. The values of δ_1, δ_2 in scenario (a) are based on the point estimate of the average treatment effect from the MISTIE II trial. We had the following goals: (i) 80% power to reject H_{00} in scenario (a); (ii) 80% power to reject H_{01} in scenario (b); (iii) strong control of familywise Type I error rate at level $\alpha^* = 0.025$. These goals were also considered by Rosenblum et al. (2016) in the context of immediately observed outcomes and not considering baseline variables. This allows us to compare the impact of delayed outcomes and adjusted estimators versus the simpler case of immediately observed outcomes considered there.

4.6.2 Data Generating Distributions used in Simulation Study

To make our simulations realistic, we mimic features in the data from the completed MISTIE II trial introduced in Section 4.2. A simple approach would be to construct simulated trials by resampling with replacement from the MISTIE II data, so that the data generating distribution is the empirical distribution of the MISTIE II data. Unfortunately, the resampling distribution does not satisfy the key feature of a randomized trial that is set by design, i.e., that treatment A is assigned independently of baseline variables. This assumption is violated in the empirical

distribution of the MISTIE II data since there are slight correlations between baseline variables and treatment assignment in the actual MISTIE II data set (as would generally be expected in any given dataset). Furthermore, since no two participants in this data have identical values of the baseline variables W, the treatment A is a deterministic function of W.

We construct data generating distributions that mimic key features of the MISTIE II data, while satisfying the randomization assumption. Specifically, we construct distributions that have similar correlations among $W, L^{(1)}, Y$ as the Phase II trial data. This is achieved by augmenting the original data set by adding, for each participant, a "twin" participant with identical baseline variables but opposite treatment assignment, whose L and Y values are generated using regression models fit to the original data, with perturbations to the outcomes Y depending on the desired treatment effect in each subpopulation. Each scenario has a different data generating distribution, based on resampling with replacement from a suitably augmented data set, as described in Section 4.9.4 of the Appendix. We assume the enrollment rate is constant over time.

4.6.3 Specific Adaptive Enrichment Design Used

We define our adaptive enrichment design by first giving the multiple testing procedure and enrollment modification rule, and then presenting the interim analysis timing and error spending functions. We use multiple testing procedure

M, described in Section 4.5. The enrollment modification rule r_k involves futility boundaries $l_{j,k}$ defined below. The following encodes our enrollment modification rule (and indicates when null hypotheses are rejected, based on multiple testing procedure M) at the interim analysis just after completion of stage $k \leq K$:

1. if $Z_{0,k} > u_{0,k}$ or $Z_{1,k} > u_{1,k}$ reject the corresponding null hypotheses and stop the trial;

2. else, if $Z_{1,k} \leq l_{1,k}$ or $k = K$ stop all enrollment;

3. else, if both subpopulations were enrolled during stage k and $Z_{2,k} \leq l_{2,k}$, stop subpopulation 2 enrollment but continue subpopulation 1 enrollment in stage $k + 1$;

4. else, if $k < K$ enroll the same subpopulations in stage $k + 1$ as in stage k.

This design strongly controls the familywise Type I error rate at level α^*. Step 2 is motivated by the clinical investigator's judgment that if the treatment benefits any subpopulation, it will very likely benefit subpopulation 1, so we should stop the entire trial for futility if $Z_{1,k}$ is below its futility boundary $l_{1,k}$. The above design (steps 1-4) is just one possible choice; our general method can be applied for any enrollment rule r_k if multiple testing procedure M is used.

It remains to define the timing of interim analyses and the error spending functions. We set the maximum number of stages $K = 5$. Alpha spending functions are from the ρ-family of Jennison and Turnbull (1999, Section 7.3) at $\rho = 2$, i.e.,

$\alpha_j(t) = c_j \min\{t^2, 1\}$ for each population $\tilde{\mathcal{S}}_j : j \in \{0, 1\}$, for nonnegative coefficients c_0, c_1 that sum to 0.025. The values of c_0, c_1, $\mathcal{I}_{0,\max}$, $\mathcal{I}_{1,\max}$, and each $l_{j,k}$ were chosen by searching over a set of candidate values to find those that minimize the average of the expected sample size over scenarios (a)-(c), under the constraint that goals (i)-(iii) are satisfied. We followed the optimization procedure from Rosenblum et al. (2016) in conducting this search. This results in $c_0 = 0.003$, $c_1 = 0.022$, $\mathcal{I}_{0,\max} = 1115$, $\mathcal{I}_{1,\max} = 795$ and the $l_{j,k}$ values given in Table 4.1. The futility boundaries $l_{j,k}$ equal 0 in most cases, with the notable exception $l_{2,3} = \infty$; this causes enrollment of subpopulation to stop at or before interim analysis 3; intuitively, this is because at interim analysis 3, sufficient information has accrued to achieve goal (i), so that further enrollment of subpopulation 2 would be counterproductive.

We use information-based monitoring times, i.e., we set the timing of interim analyses to occur when the accrued information reaches certain preset fractions of the maximum information values $\mathcal{I}_{j,\max}, j = 0, 1$ (which were preset as above based on desired values of power and Type I error under scenarios (a)-(c), and do not differ by the estimator used). The impact is that interim analyses will occur earlier in terms of calendar time (and number of participants enrolled) when using the adjusted estimator rather than the unadjusted estimator, since information accrues more quickly for the former. Each of the first three interim analyses occurs when the information accrued for the combined population reaches the equally spaced increments $1/3, 2/3, 1$ times $\mathcal{I}_{0,\max}$, i.e., each interim analysis $k \in \{1, 2, 3\}$ is

triggered when the accrued information for the combined population reaches the threshold $\mathcal{I}_{0,k} = t_{0,k}\mathcal{I}_{0,\max} = (k/3)\mathcal{I}_{0,\max}$. If subpopulation 2 is stopped before the end of stage 3, then future interim analysis timing is based on accrued information for subpopulation 1, which has equal increments for the first three stages. Since our rule r_k always stops subpopulation 2 at or before interim analysis 3, we set the 4th and 5th interim analysis times based on the information for subpopulation 1; specifically, interim analysis 4 occurs when the accrued information for subpopulation 1 reaches the midpoint between the information that accrued by the end of stage 3 ($\mathcal{I}_{1,3}$) and the maximum $\mathcal{I}_{1,\max}$.

In addition to comparing the unadjusted and adjusted estimators, we consider a modified data generating distribution where the baseline variables W and the short-term outcome $L^{(1)}$ are exogenous, i.e., independent of the treatment and outcome. This is to assess whether the adjusted estimator performs worse than the unadjusted estimator when the adjustment variables are pure noise. We denote the TMLE under this type of data generating distribution by TMLE $\text{prog}_{\emptyset}$, and denote the TMLE using the data generating distributions in Section 4.6.2 (where W and $L^{(1)}$ are prognostic) by TMLE $\text{prog}_{W,L}$.

For each stage and each estimator, Table 4.1 shows the cumulative sample sizes corresponding to the above information levels. Because information accrues more quickly when using the adjusted estimator TMLE $\text{prog}_{W,L}$, its corresponding sample sizes are smaller than those for the unadjusted estimator and for TMLE $\text{prog}_{\emptyset}$.

Table 4.2 shows the per-stage information levels $\mathcal{I}_{j,k}$ for each estimator. There are small differences between the corresponding values for each estimator due to our interim analysis times being selected from a discrete set of calendar times (that were closest to achieving the information thresholds above).

Table 4.3 gives the increments $\alpha_{j,k}$ for the unadjusted estimator, which are computed at the end of each stage k from the error spending function $\alpha_j(t) = c_j \min\{t^2, 1\}$ evaluated at $t = t_{j,k} = \mathcal{I}_{j,k}/\mathcal{I}_{j,\max}$. These values differ very slightly by scenario and by estimator due to slight differences in the covariance matrix Σ and the information levels at which analyses are conducted.

Table 4.1: Adaptive enrichment design per-stage sample sizes for scenarios (a) - (c). The cumulative sample size (Cum.S.S.) at each interim analysis has the format: number of participants with Y observed (+ number enrolled with Y not yet observed).

Interim Analysis (k)	1	2	3	4	5
			Unadjusted estimator		
Cum.S.S. Subpop. 1	104 (+24)	208 (+24)	312 (+24)	480 (+24)	648 (+0)
Cum.S.S. Subpop. 2	208 (+49)	416 (+49)	624 (+0)	624 (+0)	624 (+0)
Cum.S.S. Comb. Pop.	312 (+73)	624 (+73)	936 (+24)	1104 (+24)	1272 (+0)
			Adjusted estimator (TMLE $\text{prog}_{W,L}$)		
Cum.S.S. Subpop. 1	84 (+24)	168 (+24)	252 (+24)	382 (+24)	512 (+0)
Cum.S.S. Subpop. 2	168 (+49)	336 (+49)	504 (+0)	504 (+0)	504 (+0)
Cum.S.S. Comb. Pop.	252 (+73)	504 (+73)	756 (+24)	886 (+24)	1016 (+0)
Futility Boundary ($l_{1,k}$)	0	0	0	0	-
Futility Boundary ($l_{2,k}$)	0	0	∞	-	-

The efficacy boundaries $u_{j,k}$, which are determined by (4.1), depend on the covariance matrix Σ of the statistics under consideration. To ease the computational

Table 4.2: When using the adjusted or unadjusted estimator, information accrued at each stage for subpopulation 1, subpopulation 2, and the combined population.

	adjusted estimator					unadjusted estimator				
	stg 1	stg 2	stg 3	stg 4	stg 5	stg 1	stg 2	stg 3	stg 4	stg 5
Subpop. 1	124	250	370	558	749	126	251	376	590	795
Subpop. 2	256	524	763			249	487	739		
Comb. Pop.	389	785	1140			372	740	1115		

Table 4.3: Type I error (α) spent at each stage, for the unadjusted estimator in scenario (a). Results are very similar for the other scenarios and the adjusted estimator.

Interim Analysis (k)	1	2	3	4	5
$\alpha_{0,k}$ (for Comb. Pop.)	0.0004	0.0011	0.0016		
$\alpha_{1,k}$ (for Subpop. 1)	0.0006	0.0019	0.0029	0.0068	0.0098

Table 4.4: Efficacy boundaries for scenario (a) and unadjusted estimator. Efficacy boundaries under other scenarios and for the adjusted estimator are very similar, with all absolute differences within 0.01 of those below.

Interim Analysis (k)	1	2	3	4	5
H_{00} Efficacy Boundary ($u_{0,k}$)	3.41	3.06	2.84		
H_{01} Efficacy Boundary ($u_{1,k}$)	3.27	2.89	2.66	2.33	2.14

burden in our simulations, we precomputed an approximation to Σ using Monte Carlo simulation, and treated Σ as known, as described in Section 4.9.3 of the Appendix. The resulting boundaries $u_{j,k}$ for scenario (a) and the unadjusted estimator are given in Table 4.4. These boundaries were quite similar for both estimators and each scenario (a)-(c) (maximum absolute difference within 0.05).

4.6.4 Results: Power, Expected Sample Size, and Maximum Sample Size

Based on 50,000 simulated trials for each estimator and each scenario (a)-(c), we computed the empirical Type I error, power, and expected sample size (ESS, defined as the expected number enrolled, which includes those in the pipeline). In computing Type I error, we assume there is no early futility stopping, in order to ensure the familywise Type I error rate is at most 0.025 even when futility boundaries are ignored. In all other computations, we assume the futility boundaries are adhered to.

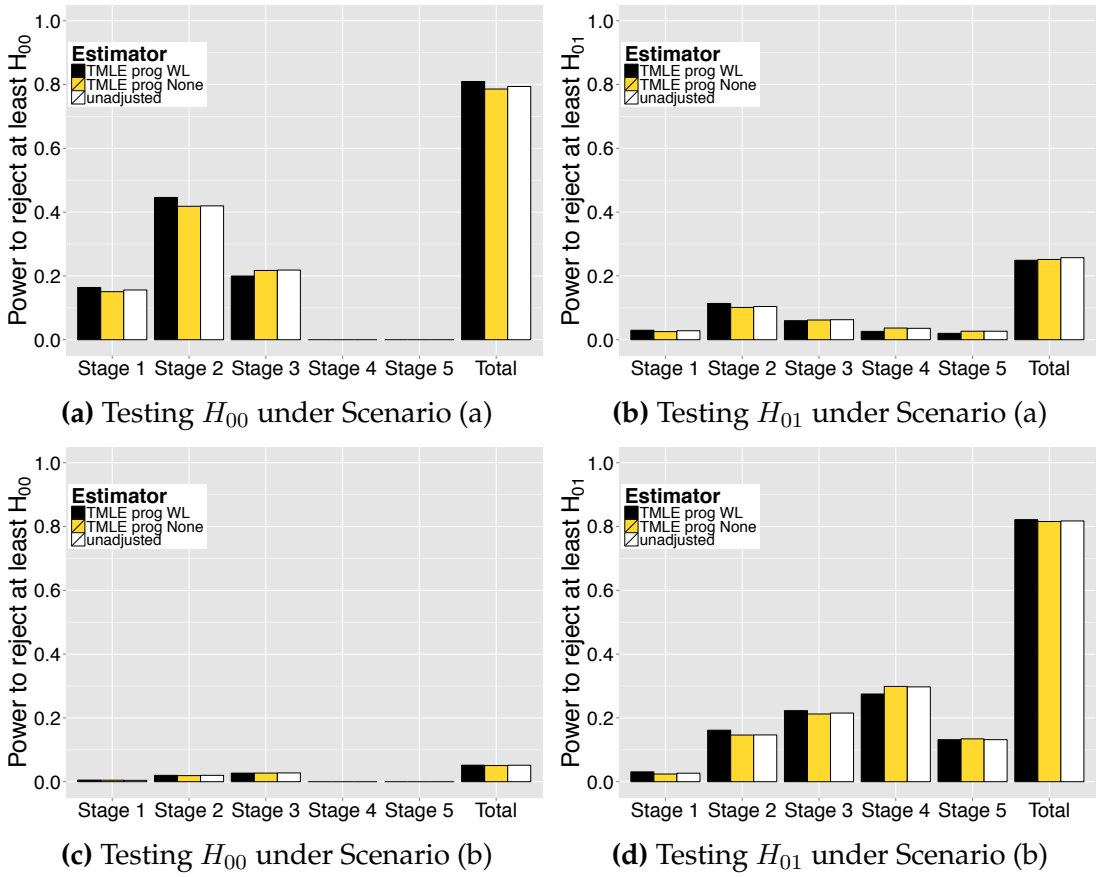
Figure 4.1 shows the probability of rejecting at least H_{00} and at least H_{01} at each interim analysis. Plots 4.1a and 4.1d display power to reject at least H_{00} under scenario (a), and power to reject at least H_{01} under scenario (b), respectively. These plots demonstrate that goals (i) and (ii) from Section 4.6.1 are approximately achieved by all estimators. Plots 4.1b and 4.1c show low power for all estimators

to reject at least H_{00} when only subpopulation 1 benefits, and to reject at least H_{01} when both subpopulations benefit. This behavior may be regarded as advantageous since it is ideal to reject only H_{00} in scenario (a) and only H_{01} in scenario (b), these corresponding precisely to the populations who benefit in each scenario, respectively.

The top half of Table 4.5 corresponds to the adaptive design from Section 4.6.3, and shows the value added from using the adjusted estimator when baseline variables and short-term outcomes are prognostic. We compare the unadjusted estimator, the adjusted estimator (denoted TMLE $\text{prog}_{W,L}$), and the adjusted estimator where W and L are set to be exogenous (denoted $\text{prog}_{\emptyset}$). Using the adjusted estimator (TMLE $\text{prog}_{W,L}$) instead of the unadjusted estimator (unadj) leads to a reduction in expected sample size of 20% under scenario (a), 19% under scenario (b), and 19% under scenario (c). Also, the maximum sample size for the design using the adjusted estimator (TMLE $\text{prog}_{W,L}$) is 20% less than that for the unadjusted estimator. In each of scenarios (a) and (b), the power of each estimator is very similar due to the information-based design using the same $\mathcal{I}_{\max}$ for each estimator. All the gains from adjusting for prognostic W and L are channeled to reducing the expected sample size. In scenario (c), the familywise Type I error rate (assuming no early stopping) is 0.025 for each estimator, as desired; early futility stopping would only decrease the Type I error. Comparing the unadjusted estimator versus TMLE $\text{prog}_{\emptyset}$ shows that when W and L provide no prognostic information, the adjusted

estimator is almost identical to the unadjusted estimator in power and expected sample size.

The design from Section 4.6.3 involves a rule for early stopping of subpopulation 2 for futility while subpopulation 1 is continued if $Z_{2,k} \leq l_{2,k}$ at interim analysis $k = 1$ or $k = 2$; we call this the adaptive enrichment feature. To show the value added by this feature, consider the same design except setting $l_{2,1} = l_{2,2} = -\infty$, which disables the ability to stop only subpopulation 2 for futility based on $Z_{2,k}$ before stage 3. We call this the non-adaptive design, and its performance in scenarios (a)-(c) is shown in the bottom half of Table 4.5. The main difference between this design and the adaptive design from Section 4.6.3 is that the former has substantially larger expected sample size in scenarios (b) and (c). This is not surprising, since it is in these scenarios when futility stopping of subpopulation 2 is especially useful. The two designs have similar power and Type I error rate in all three scenarios, and similar expected sample sizes in scenario (a). In the non-adaptive design, the adjusted estimator (TMLE $\text{prog}_{W,L}$) substantially reduces the expected sample size compared to the unadjusted estimator, just as for the adaptive design. Though it is possible to further modify the non-adaptive design to remove the rule that subpopulation 2 enrollment is always stopped at or before the end of stage 3, this would only result in increased expected sample sizes in scenarios (b) and (c), and would not help attain the goals from Section 4.6.1 since these are already met by the non-adaptive design.



(a) Testing H_{00} under Scenario (a)

(b) Testing H_{01} under Scenario (a)

(c) Testing H_{00} under Scenario (b)

(d) Testing H_{01} under Scenario (b)

Figure 4.1: Stage-wise and overall power bars comparing TMLE and unadjusted estimator. Top row corresponds to scenario (a); bottom row corresponds to scenario (b). Left column represents power to reject at least H_{00}; right column represents power to reject at least H_{01}. Black bar corresponds to adjusted estimator that leverages prognostic W and L; yellow bar corresponds to adjusted estimator with non-prognostic W and L; white bar corresponds to the unadjusted estimator.

Table 4.5: Power and expected sample size (ESS) for adaptive and non-adaptive designs. Power under scenario (a) is the probability of rejecting at least H_{00}; power under scenario (b) is the probability of rejecting at least H_{01}.

	Adaptive Design					
	Scenario (a)		Scenario (b)		Scenario (c)	
	power H_{00}	ESS	power H_{01}	ESS	Type I error	ESS
Estimator:						
unadjusted	0.79	712	0.82	795	0.025	640
TMLE $\text{prog}_{W,L}$	0.81	568	0.82	643	0.025	521
TMLE $\text{prog}_{\emptyset}$	0.79	711	0.82	794	0.025	638

	Non-Adaptive Design					
	Scenario (a)		Scenario (b)		Scenario (c)	
	power H_{00}	ESS	power H_{01}	ESS	Type I error	ESS
Estimator:						
unadjusted	0.80	718	0.82	958	0.025	729
TMLE $\text{prog}_{W,L}$	0.82	575	0.82	771	0.025	591
TMLE $\text{prog}_{\emptyset}$	0.80	718	0.81	959	0.025	727

4.7 Bias, Variance, and Mean Squared Error of Estimators

We now change focus from hypothesis testing to estimation of the average treatment effect for each subpopulation and the combined population when using the adaptive design in Section 4.6. For a given subpopulation s, consider the unadjusted and adjusted estimators evaluated at the end of the last stage in which subpopulation s gets enrolled. These estimators generally suffer from selection bias induced by early stopping rules. Our goal is to investigate whether such bias is better or worse when using the adjusted/unadjusted estimator, and when using

the adaptive design from Section 4.6.3 versus the corresponding non-adaptive design defined in 4.6.4. We consider not only the bias, but also the standard error and mean squared error of the unadjusted and adjusted estimators in scenarios (a)-(c).

We first summarize our findings. The bias, standard error, and mean squared error are very similar for the unadjusted and adjusted estimators in all cases. Also, these quantities are very similar comparing the adaptive design from Section 4.6.3 to the corresponding non-adaptive design defined in 4.6.4. This is reassuring, in that these quantities are not worse when using the adaptive design and/or the adjusted estimator (each of which provides benefits in terms of sample size reduction).

We consider the design from Section 4.6.3. Let Γ denote the covariance matrix of the estimators $\{\hat{\delta}_{1,k}, \hat{\delta}_{2,k}, \hat{\Delta}_{0,k} : k = 1, 2, 3, 4, 5\}$, which is related to the covariance Σ of the corresponding Wald statistics by $\Sigma_{ij} = \Gamma_{ij}(\Gamma_{ii}\Gamma_{jj})^{-1/2}$ for each i, j. Let $\hat{k}(s)$ denote the last stage in which subpopulation s is enrolled. This is a random quantity, since early stopping of a subpopulation depends on accrued data. For a given stagewise estimator $\hat{\delta}_{s,k}$ of δ_s, define $\delta_s^\star = \hat{\delta}_{s,\hat{k}(s)}$, i.e., the estimator evaluated at the last stage subpopulation s is enrolled. Similarly, for any composite population $\tilde{\mathcal{S}}_j$ and estimator $\hat{\Delta}_{j,k}$, define the final estimator as the following weighted combination of the final estimators for each of its component subpopulations:

$$\Delta_j^\star = \sum_{s \in \tilde{\mathcal{S}}_j} p_s \delta_s^\star \Big/ \sum_{s \in \tilde{\mathcal{S}}_j} p_s. \tag{4.2}$$

We focus on estimators of the average treatment effect for each subpopulation and for the combined population. For a given estimator (adjusted or unadjusted), design, and data generating distribution, the bias in estimating the subpopulation s average treatment effect δ_s is $E\left[\delta_s^\star\right] - \delta_s$, the standard error is $\{E[\delta_s^\star - E[\delta_s^\star]]^2\}^{1/2}$ while the mean squared error is $E[\delta_s^\star - \delta_s]^2$.

Computation is based on Monte-Carlo simulation. To make the computation feasible, instead of generating a data vector for each trial participant, we use the asymptotic approximation of the estimators $\{\hat{\delta}_{1,k}, \hat{\delta}_{2,k}, \hat{\Delta}_{0,k} : k = 1, 2, 3, 4, 5\}$ and corresponding Wald statistics $\{Z_{1,k}, Z_{2,k}, Z_{0,k}\}, k = 1, \ldots, 5$ as having a multivariate normal distributions with covariance matrices Γ and Σ computed by the method in the last paragraph of Section 4.6.3. The standard error $se_{s,k}$ of the estimator $\hat{\delta}_{s,k}$ is the square root of the corresponding element on the main diagonal of Γ.

Given a pair of treatment effects δ_1, δ_2, we generated 50,000 independent random vectors of the Wald statistics $\{Z_{1,k}, Z_{2,k}, Z_{0,k}\}, k = 1, \ldots, 5$ from the corresponding multivariate normal distribution using the rmvnorm function from the mvtnorm R package (Genz et al., 2014). For each vector, the stopping time $\hat{k}(s)$ is then determined for each subpopulation s based on the stopping rule in Section 4.6.3. Next, we set $\delta_s^\star = \hat{\delta}_{s,\hat{k}(s)} = Z_{s,\hat{k}(s)} se_{s,\hat{k}(s)}$ using the precomputed standard errors $se_{s,k}$. The estimator $\Delta_0^\star$ is set using equation (4.2). The above procedure results in 50,000 vectors of estimators $(\delta_1^\star, \delta_2^\star, \Delta_0^\star)$, from which the empirical bias, standard error and mean squared error can be computed. For example, the average

over the 50,000 draws of $(\delta_1^\star, \delta_2^\star, \Delta_0^\star)$ gives an approximation of $(E[\delta_1^\star], E[\delta_2^\star], E[\Delta_0^\star])$, which are used to approximate bias for each subpopulation and the combined population, i.e., $(E[\delta_1^\star] - \delta_1, E[\delta_2^\star] - \delta_2, E[\Delta_0^\star] - \Delta_0)$.

Table 4.6 shows that the bias, standard error and mean squared error are very similar for the adaptive design and standard design(where we will always enroll subpopulation 2 until stage 3 if no early termination of the whole trial), and for the unadjusted and adjusted estimators in scenarios (a)-(c).

Table 4.6: Approximate Bias, Standard Error (SE), and Mean Squared Error (MSE). The top half is for the unadjusted estimator and the bottom half is for the adjusted estimator (TMLE). The left side is for the adaptive design and the right side is for the non-adaptive design.

		Adaptive Design			**Non-adaptive Design**		
Estimator	Parameter	Scenario (a)	Scenario (b)	Scenario (c)	Scenario (a)	Scenario (b)	Scenario (c)
Unadjusted	δ_1	Bias=0.005 SE=0.081 MSE=0.007	Bias=0.006 SE=0.078 MSE=0.006	Bias=-0.032 SE=0.062 MSE=0.005	Bias=0.005 SE=0.081 MSE=0.007	Bias=0.006 SE=0.078 MSE=0.006	Bias=-0.032 SE=0.062 MSE=0.005
	δ_2	Bias=0.007 SE=0.056 MSE=0.003	Bias=-0.014 SE=0.052 MSE=0.003	Bias=-0.008 SE=0.058 MSE=0.003	Bias=0.009 SE=0.051 MSE=0.003	Bias=0.000 SE=0.043 MSE=0.002	Bias=0.000 SE=0.053 MSE=0.003
	Δ_0	Bias=0.006 SE=0.047 MSE=0.002	Bias=-0.008 SE=0.043 MSE=0.002	Bias=-0.016 SE=0.042 MSE=0.002	Bias=0.008 SE=0.044 MSE=0.002	Bias=0.002 SE=0.039 MSE=0.002	Bias=-0.011 SE=0.041 MSE=0.002
Adjusted (TMLE)	δ_1	Bias=0.004 SE=0.082 MSE=0.007	Bias=0.006 SE=0.077 MSE=0.006	Bias=-0.032 SE=0.063 MSE=0.005	Bias=0.004 SE=0.082 MSE=0.007	Bias=0.006 SE=0.077 MSE=0.006	Bias=-0.032 SE=0.063 MSE=0.005
	δ_2	Bias=0.007 SE=0.054 MSE=0.003	Bias=-0.014 SE=0.052 MSE=0.003	Bias=-0.008 SE=0.057 MSE=0.003	Bias=0.009 SE=0.050 MSE=0.003	Bias=0.000 SE=0.043 MSE=0.002	Bias=0.000 SE=0.052 MSE=0.003
	Δ_0	Bias=0.006 SE=0.047 MSE=0.002	Bias=-0.007 SE=0.043 MSE=0.002	Bias=-0.016 SE=0.042 MSE=0.002	Bias=0.007 SE=0.045 MSE=0.002	Bias=0.002 SE=0.039 MSE=0.002	Bias=-0.011 SE=0.041 MSE=0.002

4.8 Remarks

Alternative methods exist for covariate adjustment in our longitudinal setting, e.g., the estimators of Lu and Tsiatis (2011); Rotnitzky et al. (2012); Gruber and van der Laan (2012). These estimators have enhanced efficiency properties, but to the best of our knowledge there is not currently an R package implementing any of these methods that incorporates both baseline variables and short-term outcomes. It is an area of future work to apply these advanced methods in the context of adaptive enrichment designs.

We assumed that the only cause of missing data was administrative censoring due to some participants not yet having their final outcomes observed. In Section 4.9.2 of the Appendix, we describe how to incorporate additional right censoring due to loss to follow-up, under the assumption of censoring at random, i.e., the sequential randomization assumption in (van der Laan and Gruber, 2012) stating that censoring is conditionally independent of the potential outcomes given the past observed data.

We focused on outcomes measured a fixed duration from enrollment. Our adaptive design framework in Section 4.5 can also be applied for survival times, e.g., by using a modified TMLE or the estimators of Lu and Tsiatis (2011); Parast et al. (2014); Zhang (2015).

When a decision is made to stop a subpopulation or the entire trial early for efficacy, the corresponding null hypothesis is immediately rejected without waiting

for pipeline patients to complete the trial. It is possible to improve efficiency by waiting until pipeline patients complete the trial, and using a modified test that takes their outcomes into account, e.g., by extending the approach of Hampson and Jennison (2013) that was developed for standard, group sequential designs and a single population. It is an open problem to extend this method to our setting where we desire strong control of the familywise Type I error rate, since this is not guaranteed if we directly apply their method in our context (which their method was not designed for).

We stated in Section 4.4 that the covariance matrix Σ does not generally have the canonical structure from (Scharfstein et al., 1997; Jennison and Turnbull, 1999) in the case where $\hat{\Delta}_{j,k}$ is the adjusted estimator. This does not contradict the main result of Scharfstein et al. (1997), since the adjusted estimator is generally not globally, semiparametric efficient (only locally efficient); in general, no globally, semiparametric efficient estimator exists for our problem, unless one makes model assumptions on the outcome distributions (which we avoid since these cause biased and difficult to interpret estimates when the model assumptions fail to hold). The above issue does not pose any problem for our framework in Section 4.5, since the error spending function approach can be applied without requiring the canonical structure of Scharfstein et al. (1997); Jennison and Turnbull (1999).

We focused on the case of two subpopulations of interest. The general framework in Section 4.5 can be applied to any number of subpopulations and composite

populations. However, as the number of such populations increases, the required sample size to achieve high power for each population (while maintaining strong control of the familywise Type I error rate) will also increase. It is an open problem to determine how many populations can be accommodated before sample size becomes prohibitively large.

4.9 Appendix

4.9.1 Regularity Conditions for Q and $Q^{(W)}$ and Asymptotic Results

We define our asymptotic framework, and then show that the unadjusted and adjusted estimators are consistent and asymptotically normal. Consider any distributions Q and $Q^{(W)}$ that satisfy the regularity conditions in Theorem A5 in Appendix A18 of van der Laan and Rose (2011). These conditions imply the TMLE that we used, described in Section 4.9.2 of the Appendix below, is consistent and asymptotically normal, under independent, identically distributed sampling of participants. The sampling in our designs is slightly different, but the same results hold in our setting, as described below.

We use the notation $L^{(T+1)}$ to represent the final outcome Y. Let $N_{s,k}^{(t)}$ denote the cumulative number of participants enrolled from subpopulation s by the end of

stage k who have (at least) $L^{(t)}$ observed, assuming no early stopping, i.e., $N_{s,k}^{(t)} = \sum_i C_{i,k}^{(t)} 1[S_i = s]$. Similarly, let $\tilde{N}_{j,k}^{(t)}$ denote the cumulative number of participants enrolled from composite population $\tilde{\mathcal{S}}_j$ by the end of stage k who have (at least) $L^{(t)}$ observed, assuming no early stopping, i.e., $\tilde{N}_{j,k}^{(t)} = \sum_i C_{i,k}^{(t)} 1[S_i \in \tilde{\mathcal{S}}_j]$. Define $N_{s,k} = N_{s,k}^{(0)}$, i.e., the cumulative number enrolled from subpopulation s by the end of stage k, assuming no early stopping. Similarly, let $\tilde{N}_{j,k} = \tilde{N}_{j,k}^{(0)}$, i.e., the the cumulative number enrolled from composite population $\tilde{\mathcal{S}}_j$ by the end of stage k, assuming no early stopping. Let $C_{i,k} = \{C_{i,k}^{(t)}\}_{t=0}^{T+1}$.

Let e denote the combined population enrollment rate in number of participants per day, which we assume to be constant over time. By the assumption from Section 4.3 of sampling proportional to subpopulation size, the enrollment rate for subpopulation s is ep_s. The maximum trial duration is the time from the start of enrollment to the time the last participant has his/her final outcome Y observed if there is no early stopping, which equals

$$\mathcal{D} = \max_{s \in \{1,\ldots,m\}} N_{s,\max}/(ep_s) + d_Y. \tag{4.3}$$

(Recall d_Y is the delay time from enrollment to observation of Y.) The maximum total sample size is denoted by $N = \sum_{s \in \{1,\ldots,m\}} N_{s,\max}$.

In our asymptotic framework, we fix the maximum trial duration, the delay times $d_1, \ldots, d_T, d_Y$, and the interim analysis times. These are in terms of calendar

time, but can equivalently be expressed in terms of information time. We let the maximum total sample size N go to infinity such that the cumulative sample size proportions $\{N_{s,k}/N\}_{s\leq m,k\leq K}$ converge to a fixed set of nonnegative proportionality constants $\{q_{s,k}\}_{s\leq m,k\leq K}$ that satisfy $q_{s,k} \leq q_{s,k+1}$ for each $k < K$ and $\sum_s q_{s,K} = 1$. Let $\tilde{q}_{j,k} = \sum_{s\in\tilde{\mathcal{S}}_j} q_{s,k}$. Since we fix the maximum trial duration $\mathcal{D}$ and delay times $d_1, \ldots, d_T, d_Y$, this implies by (4.3) that the combined population enrollment rate e goes to infinity and is (in the limit) proportional to N, i.e.,

$$\lim_{N\to\infty} \frac{e}{N} = \lim_{N\to\infty} \frac{1}{N} \max_{s\in\{1,\ldots,m\}} N_{s,\max}/\{p_s(\mathcal{D}-d_Y)\} = \max_{s\in\{1,\ldots,m\}} q_{s,K}/\{p_s(\mathcal{D}-d_Y)\}, \quad (4.4)$$

where the rightmost term is a nonnegative constant.

The number of pipeline participants in subpopulation s at interim analysis $k < K$, assuming enrollment has not stopped for subpopulation s, is $ep_s d_Y$; therefore, by (4.4), the proportion of enrolled participants from subpopulation s that are in the pipeline at interim analysis k is

$$\lim_{N\to\infty} \frac{ep_s d_Y}{N_{s,k}} = \lim_{N\to\infty} \frac{d_Y ep_s/N_{\max}}{N_{s,k}/N_{\max}} = \frac{d_Y}{\mathcal{D}-d_Y} \frac{\max_{s'\in\{1,\ldots,m\}} q_{s',K}}{q_{s,k}}.$$

This shows that the proportion of enrolled participants from subpopulation s that are in the pipeline at interim analysis k converges to a nonnegative constant (that depends on s and k). This is a desirable feature of our asymptotic framework, since it allows us to consider the impact, in the limit, of a constant fraction of par-

ticipants in the pipeline at each interim analysis. In contrast, if this proportion had converged to zero then the asymptotic framework would not reflect the impact of pipeline participants (since their impact would disappear asymptotically). We will see that the pipeline participants can improve precision of the adjusted estimator, if baseline variables or short-term outcomes are prognostic of the final outcome.

Our asymptotic framework has the following similarities with that of Scharfstein et al. (1997, Section 2): data are collected over a fixed time interval (that does not change with sample size); each participant's data is a longitudinal process; sample size goes to infinity, which implies the average number enrolled per unit of time goes to infinity and that the expected proportion enrolled at each interim analysis k converges to a constant (that depends on k); in the special case of fixed delay times, the expected proportion in the pipeline at interim analysis k converges to a constant (that depends on k). The main differences between these frameworks are the following: we assume a constant enrollment rate and fixed delay times while they consider a random enrollment process and participant data can be measured at variable times (or in continuous time); they assume the existence of a globally, semiparametric efficient estimator, while we do not. The main reason we focus on a constant enrollment rate is that it makes computations faster in our simulations, since for any $Q, Q^{(W)}$ and any information-based monitoring times, the stage-wise cumulative sample sizes $N_{s,k}^{(t)}$ are then fixed. It is an area of future work to let these sample sizes be variable (depending on randomness in the enrollment procedure),

and to do all analyses conditional on the observed enrollment times. The framework of Scharfstein et al. (1997, Section 2) also has the important advantage that it allows for changes to the population distribution over time; however, it requires existence of a globally semiparametric efficient estimator, which is generally not available for our problem setting without making parametric model assumptions or smoothness assumptions on the model for $Q, Q^{(W)}$, which we do not do.

In order to illustrate the main idea behind our results, we first consider the special case where the following hold: there is no early stopping, i.e., each subpopulation s is enrolled until $N_{s,\max}$ is reached; $T = 0$ (no short-term outcome); and the only cause of missing data is administrative censoring (as defined in Section 4.3.5). Even in this special case, Wald statistics based on the adjusted estimator (TMLE) do not generally have the canonical covariance matrix of Scharfstein et al. (1997); Jennison and Turnbull (1999). In this special case, the censoring times $C_{i,k}$ are fixed (non-random) and independent of the participant data $\{D_i\}_{i=1}^N$. It also follows by the assumptions in Section 4.3.3 that the data $\{D_i\}_{i=1}^N$ are independent, identically distributed vectors $D_i = (S_i, W_i, A_i, Y_i)$ drawn from the data generating distribution $\tilde{Q} = (Q, Q^{(W)})$ (and $P(A = 1|W, S) = 1/2$ by the randomization assumption). We let D denote a generic vector with components (S, W, A, Y) having this distribution. We use the TMLE $\hat{\delta}_{s,k}$ defined in Section 4.9.2 that uses the known quantities $P(A|W, S) = 1/2$ and the known fractions of total participants $i \in \{1, \ldots, N\}$ who have $C_{i,k}^{(t)} = 1$, i.e., the fraction $N_{s,k}^{(t)}/N$. Let $P^*(Y = 1|A, W, S)$

denote the limit in probability of the working logistic regression model for the corresponding probability; this model will generally be misspecified, and P^* denotes its limit which may differ from the true conditional probability. Under the aforementioned regularity conditions, $\hat{\delta}_{s,k}$ is consistent for the average treatment effect δ_s and is asymptotically normal, i.e., (where $o_p(1)$ represents a quantity that converges to 0 in probability as N goes to infinity)

$$\sqrt{N}(\hat{\delta}_{s,k} - \delta_s) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} IF_{s,k}(D_i) + o_p(1), \tag{4.5}$$

where the influence function $IF_{s,k}$ for the estimator $\hat{\delta}_{s,k}$ is defined as

$$\begin{aligned} IF_{s,k}(D_i) \;=\; & \frac{C_{i,k}^{(1)} 1\{S_i = s\}}{N_{s,k}^{(1)}/N} \frac{2A_i - 1}{1/2} \left\{ Y_i - P^*(Y = 1 | A_i, W_i, S_i) \right\} \qquad (4.6) \\ & + \frac{C_{i,k}^{(0)} 1\{S_i = s\}}{N_{s,k}^{(0)}/N} \left\{ P^*(Y = 1 | A = 1, W_i, S_i) - P^*(Y = 1 | A = 0, W_i, S_i) - \delta_s \right\}. \qquad (4.7) \end{aligned}$$

The influence function $IF_{s,k}(D_i)$ depends on $\tilde{Q}$ and the fixed (non-random) set of censoring times $C_k = \{C_{i,k}\}_{i=1}^N$, but we suppress this in our notation for conciseness. The above asymptotic distribution follows from Theorem A5 in Appendix A18 of van der Laan and Rose (2011), except using the empirical process result for non-identically distributed (but independent) random variables of Alexander (1984) that handles a fixed sequence of censoring times rather than random, in-

dependent, identically distributed censoring times; the results are analogous by the assumption above that the censoring is due only to administrative censoring that is independent of the participant data D_i. The form of the influence function $IF_{s,k}(D_i)$ is similar to the form in van der Laan and Gruber (2012) except that we use the known value $1/2$ in place of $P(A = A_i|W_i, S_i)$ and the known censoring frequency $N_{s,k}^{(t)}/N$ in place of $P(C_{i,k}^{(t)} = 1)$ for each $t \in \{0, 1\}$.

Define the following components from (4.6) and (4.7), respectively:

$$\begin{aligned} H_s^{(1)}(D) &= \frac{(2A-1)}{1/2}\left\{Y - P^*(Y = 1|A, W, S)\right\}; \\ H_s^{(0)}(D) &= \left[P^*(Y = 1|A = 1, W, S) - P^*(Y = 1|A = 0, W, S) - \delta_s\right]. \end{aligned}$$

Let Var_s and Cov_s denote variance conditional on $S = s$ and covariance conditional on $S = s$, respectively.

It follows from (4.5) that for any stages $k \leq k' \leq K$ the asymptotic covariance

of $\hat{\delta}_{s,k}$ and $\hat{\delta}_{s,k'}$ has the following form:

$$
\begin{aligned}
& \text{acov}(\hat{\delta}_{s,k}, \hat{\delta}_{s,k'}) \\
= & \lim_{N\to\infty} \text{Cov}\left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} IF_{s,k}(D_i), \frac{1}{\sqrt{N}} \sum_{i'=1}^{N} IF_{s,k'}(D_{i'}) \right\} \\
= & \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \sum_{i'=1}^{N} \text{Cov}\left\{ IF_{s,k}(D_i), IF_{s,k'}(D_{i'}) \right\} \\
= & \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \text{Cov}\left\{ IF_{s,k}(D_i), IF_{s,k'}(D_i) \right\} \\
= & \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \text{Cov}\left\{ \frac{C_{i,k}^{(1)} 1\{S_i = s\}}{N_{s,k}^{(1)}/N} H_s^{(1)}(D_i) + \frac{C_{i,k}^{(0)} 1\{S_i = s\}}{N_{s,k}^{(0)}/N} H_s^{(0)}(D_i), \right. \\
& \qquad \left. \frac{C_{i,k'}^{(1)} 1\{S_i = s\}}{N_{s,k'}^{(1)}/N} H_{s,k'}^{(1)}(D_i) + \frac{C_{i,k'}^{(0)} 1\{S_i = s\}}{N_{s,k'}^{(0)}/N} H_{s,k'}^{(0)}(D_i) \right\} \\
= & \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \left[N_{s,k}^{(0)} \text{Cov}_s \left\{ \frac{H_s^{(0)}(D)}{N_{s,k}^{(0)}/N}, \frac{H_s^{(0)}(D)}{N_{s,k'}^{(0)}/N} \right\} + N_{s,k}^{(1)} \text{Cov}_s \left\{ \frac{H_s^{(1)}(D)}{N_{s,k}^{(1)}/N}, \frac{H_s^{(0)}(D)}{N_{s,k'}^{(0)}/N} \right\} \right. \\
& \qquad \left. + N_{s,k'}^{(1)} \text{Cov}_s \left\{ \frac{H_s^{(0)}(D)}{N_{s,k}^{(0)}/N}, \frac{H_s^{(1)}(D)}{N_{s,k'}^{(1)}/N} \right\} + N_{s,k}^{(1)} \text{Cov}_s \left\{ \frac{H_s^{(1)}(D)}{N_{s,k}^{(1)}/N}, \frac{H_s^{(1)}(D)}{N_{s,k'}^{(1)}/N} \right\} \right] \\
= & \lim_{N\to\infty} \sum_{i=1}^{N} \left[\frac{N}{N_{s,k'}^{(0)}} \text{Var}_s \left\{ H_s^{(0)}(D) \right\} + \left(\frac{N}{N_{s,k'}^{(0)}} + \frac{N}{N_{s,k}^{(0)}} \right) \text{Cov}_s \left\{ H_s^{(0)}(D), H_s^{(1)}(D) \right\} \right. \\
& \qquad \left. + \frac{N}{N_{s,k'}^{(1)}} \text{Var}_s \left\{ H_s^{(1)}(D) \right\} \right] \\
= & \frac{1}{q_{s,k'}^{(1)}} \text{Var}_s \left\{ H_s^{(1)}(D) \right\} + \left\{ \frac{1}{q_{s,k}^{(0)}} + \frac{1}{q_{s,k'}^{(0)}} \right\} \text{Cov}_s \left\{ H_s^{(1)}(D), H_s^{(0)}(D) \right\} + \frac{1}{q_{s,k'}^{(0)}} \text{Var}_s \left\{ H_s^{(0)}(D) \right\}.
\end{aligned}
\tag{4.8}
$$

An important feature of (4.8) is that the covariance term does not generally equal 0 if the limit distribution $P^*(Y = 1|A, W, S)$ of the logistic regression working model

does not equal the true distribution $P(Y = 1|A, W, S)$, which we expect will generally occur in practice due to misspecification of the outcome regression model. As we discuss below, this is the reason that the Wald statistics derived from the TMLE estimator do not have the canonical covariance structure of Scharfstein et al. (1997); Jennison and Turnbull (1999).

We now give analogous results as above, but for estimators of Δ_j rather than of δ_s.

$$\sqrt{N}(\hat{\Delta}_{j,k} - \Delta_j) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \overline{IF}_{j,k}(D_i) + o_p(1), \tag{4.9}$$

where the influence function $\overline{IF}_{j,k}$ for the estimator $\hat{\Delta}_{j,k}$ is defined as

$$\begin{aligned} \overline{IF}_{j,k}(D_i) &= \frac{C_{i,k}^{(1)} 1\{S_i \in \tilde{\mathcal{S}}_j\}}{\tilde{N}_{s,k}^{(1)}/N} \frac{2A_i - 1}{1/2} \left\{Y_i - P^*(Y = 1|A_i, W_i, S_i)\right\} \qquad (4.10) \\ &\quad + \frac{C_{i,k}^{(0)} 1\{S_i \in \tilde{\mathcal{S}}_j\}}{\tilde{N}_{s,k}^{(0)}/N} \left\{P^*(Y = 1|A = 1, W_i, S_i) - P^*(Y = 1|A = 0, W_i, S_i) - \delta_s\right\}. \qquad (4.11) \end{aligned}$$

Define the following components from (4.10) and (4.11), respectively:

$$\begin{aligned} \tilde{H}_j^{(1)}(D) &= \frac{(2A - 1)}{1/2} \left\{Y - P^*(Y = 1|A, W, S)\right\}; \\ \tilde{H}_j^{(0)}(D) &= \left[P^*(Y = 1|A = 1, W, S) - P^*(Y = 1|A = 0, W, S) - \Delta_j\right]. \end{aligned}$$

Let Var_j and Cov_j denote variance conditional on $S \in \tilde{\mathcal{S}}_j$ and covariance conditional on $S \in \tilde{\mathcal{S}}_j$, respectively. By a similar derivation as above, for any stages

$k \leq k' \leq K$ the asymptotic covariance of $\hat{\Delta}_{j,k}$ and $\hat{\Delta}_{j,k'}$ equals

$$\begin{aligned} & \text{acov}(\hat{\Delta}_{j,k}, \hat{\Delta}_{j,k'}) \\ = & \lim_{N\to\infty} \text{Cov}\left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \overline{IF}_{j,k}(D_i), \frac{1}{\sqrt{N}} \sum_{i'=1}^{N} \overline{IF}_{j,k'}(D_{i'}) \right\} \\ = & \frac{1}{\tilde{q}^{(1)}_{j,k'}} \text{Var}_j \left\{ \tilde{H}^{(1)}_j(D) \right\} + \left\{ \frac{1}{\tilde{q}^{(0)}_{j,k}} + \frac{1}{\tilde{q}^{(0)}_{j,k'}} \right\} \text{Cov}_j \left\{ \tilde{H}^{(1)}_j(D), \tilde{H}^{(0)}_j(D) \right\} + \frac{1}{\tilde{q}^{(0)}_{j,k'}} \text{Var}_j \left\{ \tilde{H}^{(0)}_j(D) \right\}. \end{aligned} \tag{4.12}$$

Define $\text{avar}(\hat{\Delta}_{j,k}) = \text{acov}(\hat{\Delta}_{j,k}, \hat{\Delta}_{j,k})$, and the asymptotic information

$$\mathcal{I}^*_{j,k} = \lim_{N\to\infty} \mathcal{I}_{j,k}/N = \lim_{N\to\infty} \left\{ N\text{Var}(\hat{\Delta}_{j,k}) \right\}^{-1} = \left\{ \text{avar}(\hat{\Delta}_{j,k}) \right\}^{-1}$$

It follows from (4.9) that for $Z_{j,k}$ the Wald statistic corresponding to $\tilde{\Delta}_{j,k}$, we have

$$\sqrt{N}\left(\left\{\mathcal{I}^*_{j,1}\right\}^{1/2} Z_{j,1}, \ldots, \left\{\mathcal{I}^*_{j,K}\right\}^{1/2} Z_{j,K} \right) = \sqrt{N}\left(\mathcal{I}^*_{j,1}(\hat{\Delta}_{j,1} - \Delta_j), \ldots, \mathcal{I}^*_{j,K}(\hat{\Delta}_{j,K} - \Delta_j) \right)$$

converges to a multivariate normal distribution with zero mean vector and covariance matrix with (k, k') component equal to

$$\text{acov}(\hat{\Delta}_{j,k}, \hat{\Delta}_{j,k'}) \left\{ \text{avar}(\hat{\Delta}_{j,k}) \text{avar}(\hat{\Delta}_{j,k'}) \right\}^{-1}, \tag{4.13}$$

which does not generally equal $\mathcal{I}^*_{j,k}$ due to the covariance term in (4.12) which

depends on k as well as k'. In contrast, if that covariance term were equal to 0, then we would have acov$(\hat{\Delta}_{j,k}, \hat{\Delta}_{j,k'}) = $ avar$(\hat{\Delta}_{j,k'})$ which implies (4.13) equals $\mathcal{I}^*_{j,k}$; this means that the asymptotic covariance matrix of the (scaled) Wald statistics would have an independent increments structure as in Scharfstein et al. (1997, Theorem 1) and Jennison and Turnbull (1999, Chapter 3). The impact is that multiple testing procedures based on the p-value combination approach may not be applicable, since they require independent increments or the more general p-clud property (that under the null hypothesis, the conditional distribution of each stage k p-value given all prior stage data stochastically dominates the uniform distribution on $[0, 1]$), neither of which are guaranteed to hold in our setting when using the adjusted estimator.

The covariance term in (4.12) does not generally equal 0 in our setting if the logistic regression model used in the TMLE for $P(Y = 1|A, W, S = s)$ (called the outcome regression model) is misspecified. This is because the limit distribution $P^*(Y = 1|A, W, S)$ of such a misspecified model will not equal the true distribution $P(Y = 1|A, W, S)$, and so $Y - P^*(Y = 1|A, W, S)$ will generally not be orthogonal to all functions of A, W, S. We expect the outcome regression model to be at least somewhat misspecified in practice for W continuous valued with a few or more components.

In contrast, if the outcome regression model were correctly specified, then $P^*(Y = 1|A, W, S) = P(Y = 1|A, W, S)$ and so $Y - P^*(Y = 1|A, W, S)$ would be orthogo-

nal to all functions of A, W, S; then the covariance term in (4.8) would equal 0. The TMLE is semiparametric efficient at a given distribution $\tilde{Q}$ if and only if the limit P^* of the outcome regression model equals the true distribution of Y given A, W, S; that is why the corresponding covariance term disappears in the setting of Scharfstein et al. (1997) (where a semiparametric efficient estimator is assumed) resulting in the canonical distribution in their Theorem 1. We also note that in the special case where the outcome regression working model is a linear model, then the covariance term equals 0; however, for other working models (including logistic regression) the covariance term will not disappear.

Above we considered a special case where $T = 0$, there is no early stopping, and the only cause of missing data is administrative censoring. We now relax these constraints one by one. Consider the case of arbitrary $T > 0$. The results above are similar, except that the influence function for $\hat{\delta}_{s,k}$ has additional components, as follows, where we use the notation $\bar{L}^{(t)} = (L^{(t)}, \ldots, L^{(1)})$ and $\bar{L}^{(0)} = \emptyset$:

$$
\begin{aligned}
& IF_{s,k}^{(T)}(D_i) \\
= \; & \frac{C_{i,k}^{(T+1)} 1\{S_i = s\}}{N_{s,k}^{(T+1)}/N} \frac{2A_i - 1}{1/2} \left\{ Y_i - P^*(Y = 1 | \bar{L}^{(t)}, A_i, W_i, S_i) \right\} \\
& + \sum_{t=1}^{T} \frac{C_{i,k}^{(t)} 1\{S_i = s\}}{N_{s,k}^{(t)}/N} \frac{2A_i - 1}{1/2} \left\{ P^*(Y = 1 | \bar{L}^{(t)}, A_i, W_i, S_i) - P^*(Y = 1 | \bar{L}^{(t-1)}, A_i, W_i, S_i) \right\} \\
& + \frac{C_{i,k}^{(0)} 1\{S_i = s\}}{N_{s,k}^{(0)}/N} \left\{ P^*(Y = 1 | A = 1, W_i, S_i) - P^*(Y = 1 | A = 0, W_i, S_i) - \delta_s \right\}. \qquad (4.14)
\end{aligned}
$$

The above holds since in the case only administrative censoring and no early stopping, the censoring times $C_{i,k}^{(t)}$ are fixed (non-random) and D_i are independent, identically distributed, by the assumptions above. An analogous generalization of the influence function for $\Delta_{j,k}$ holds. The covariance formulas (4.8) and (4.12) have additional components corresponding to the additional terms in (4.14).

Consider the case where there is early stopping according to a prespecified rule r_k. Then the above asymptotic linearity result (4.5) holds only up through stages in which subpopulation s is enrolled. Similarly, the above asymptotic linearity results (4.9) and (4.13) hold only up through stages in which every subpopulation $s \in \tilde{\mathcal{S}}_j$ is enrolled. If enrollment is stopped early for any subpopulation, the corresponding estimators and Wald statistics involving that subpoplation are no longer used in the multiple testing procedures in Section 4.5, by construction. Therefore, asymptotically, the multiple testing procedure has familywise Type I error rate equal to that computed by assuming the statistics $Z_{j,k}$ have the asymptotic, multivariate normal distribution with zero mean vector and covariance matrix (4.13).

The form of the influence functions given above assumed that we use the TMLE $\hat{\delta}_{s,k}$ defined in Section 4.9.2 that uses the known quantities $P(A|W,S) = 1/2$ and the known fractions of total participants $i \in \{1,\ldots,N\}$ who have $C_{i,k}^{(t)} = 1$, i.e., the fraction $N_{s,k}^{(t)}/N$. If $P(A|W,S)$ and $P(C_{i,k}^{(t)} = 1|\bar{L}^{(t-1)}, C_{i,k}^{(t)} = 1, A, W, S)$ are instead estimated using logistic regression models (which is what we used for the TMLE in our simulations), then each influence function is as above except one subtracts

its projection on the space of nuisance scores corresponding to these models, as described in (van der Laan and Gruber, 2012, Section 4). This leads to unchanged or reduced asymptotic variance, and may also affect the asymptotic covariance. The nonparametric bootstrap can still be used as a consistent estimator of the covariance matrix of the adjusted estimator $\hat{\Delta}_{j,k}$ (and of the corresponding Wald statistics).

Last, we consider the case where data are missing not only due to administrative censoring but also due to drop-out. Under the assumption that drop-out is a random, right-censoring process satisfying the censoring at random assumption, i.e., the sequential randomization assumption in (van der Laan and Gruber, 2012), the above results can be generalized. Additional missing outcomes due to drop-out will generally reduce the information available at each interim analysis. Also, unless outcomes are assumed to be missing completely at random (i.e., independent of D), consistency of the adjusted and unadjusted estimators will generally require additional assumptions such as that the outcome regression model or censoring model is correctly specified. This reflects the double robustness property of the TMLE, which is described in (van der Laan and Gruber, 2012).

4.9.2 TMLE Estimator

We give an overview of the targeted maximum likelihood estimator of van der Laan and Gruber (2012) that is implemented in the R package **ltmle** (Schwab et al.,

2014), and describe the details of our implementation of this estimator (called the adjusted estimator in the main paper); this estimator combines features of the general targeted maximum likelihood template of van der Laan and Rubin (2006) with the sequential regression approach of (Robins, 2000; Bang and Robins, 2005). We focus on the setting with a single short-term outcome, i.e., $T = 1$, which was used in our simulation studies; the more general case of $T > 1$ is described by van der Laan and Gruber (2012). Assume the longitudinal data structure $D = (S, W, A, L, Y)$, where S is the subpopulation, W is the vector of baseline variables, A is the treatment assignment, $L = L^{(1)}$ is the short-term outcome, and Y is the final outcome. At each interim analysis $k < K$, only a subset of those enrolled have the short-term outcome observed, and a further subset have the final outcome observed.

Consider any interim analysis $k \leq K$. Let C_L denote the indicator that L is observed and let C_Y denote the indicator that Y is observed. (Formally, for each participant i, we define $C_{L,i} = C_{i,k}^{(1)}$ and $C_{Y,i} = C_{i,k}^{(2)}$.) We first give the initial working models used in the algorithm for computing TMLE, and then give the general TMLE algorithm.

We describe the TMLE estimator of $E(Y|A = a, S = s)$ for each $a \in \{0, 1\}, s \in \{1, 2\}$ using the method of van der Laan and Gruber (2012). We consider two baseline variables $W = (W_1, W_2)$. For each study arm $a \in \{0, 1\}$ and subpopulation $s \in \{1, \ldots, m\}$, define the following logistic regression working models (where

$\text{logit}(x) = \log\{x/(1-x)\}$):

$$
\begin{aligned}
P(Y=1|C_Y=1,L,C_L=1,A=a,W,S=s) &= \text{logit}^{-1}\left(\beta_0^{a,s}+\beta_1^{a,s}W_1+\beta_2^{a,s}W_2+\beta_3^{a,s}L\right), \\
P(C_Y=1|L,C_L=1,A=a,W,S=s) &= \text{logit}^{-1}\left(\eta_0^{a,s}+\eta_1^{a,s}W_1+\eta_2^{a,s}W_2+\eta_3^{a,s}L\right), \\
P(C_L=1|A=a,W,S=s) &= \text{logit}^{-1}\left(\zeta_0^{a,s}+\zeta_1^{a,s}W_1+\zeta_2^{a,s}W_2\right), \\
P(A=a|W,S=s) &= \text{logit}^{-1}\left(\xi_0^{a,s}+\xi_1^{a,s}W_1+\xi_2^{a,s}W_2\right).
\end{aligned}
$$

The following is the TMLE algorithm from van der Laan and Gruber (2012) applied to our setting:

Step 1: For each of the above logistic regression working models, fit it using maximum likelihood estimation using participants who satisfy the corresponding condition. Specifically, the first model is fit using the participants with $C_Y = C_L = 1, A = a, S = s$; the second uses participants with $C_L = 1, A = a, S = s$; the third uses participants with $A = a, S = s$; and the fourth uses participants with $S = s$. Let $\hat{\mathbb{P}}$ denote the predicted probability corresponding to each model fit above, e.g.,

$$
\begin{aligned}
&\hat{\mathbb{P}}(Y=1|C_Y=1,L,C_L=1,A=a,W,S=s) = \\
&\text{logit}^{-1}\left(\hat{\beta}_0^{a,s}+\hat{\beta}_1^{a,s}W_1+\hat{\beta}_2^{a,s}W_2+\hat{\beta}_3^{a,s}L\right),
\end{aligned}
$$

where $\hat{\beta}_l^{a,s}$ denotes the estimated coefficient $\beta_l^{a,s}$.

Step 2: Using only participants with $A = a, S = s, C_L = C_Y = 1$, fit the logistic regression model (called the updated model) $m\left(L, W, \epsilon_Y\right)$ for $P(Y = 1|C_Y = 1, L, C_L = 1, A = a, W, S = s)$ with single covariate

$$\left\{\hat{\mathbb{P}}\left(C_Y = 1 \mid L, C_L = 1, A = a, W, S = s\right)\hat{\mathbb{P}}\left(C_L = 1 \mid A = a, W, S = s\right)\hat{\mathbb{P}}\left(A = a \mid W, S = s\right)\right\}^{-1},$$

and using the logit of $\hat{\mathbb{P}}(Y = 1|C_Y = 1, L, C_L = 1, A = a, W, S = s)$ from the step 1 as offset in the linear part of the model m. Denote the fitted regression coefficient as $\hat{\epsilon}_Y$ and the predicted values based on this updated model fit (which includes the offset) by $m\left(L, W, \hat{\epsilon}_Y\right)$.

Step 3: Define the new variable $H = m\left(L, W, \hat{\epsilon}_Y\right)$ for each participant with $C_L = 1, A = a, S = s$. Fit the initial logistic regression model $m'_{init}\left(W, \gamma_{init}\right)$ for $E\left(H \mid C_L = 1, A = a, W, S = s\right)$ using participants with $C_L = 1, A = a, S = s$. Next, analogous to step 2, fit the logistic regression model $m'\left(W, \epsilon_H\right)$ for $E\left(H \mid C_L = 1, A = a, W, S = s\right)$ using this same set of participants, but with single covariate $\left\{\hat{\mathbb{P}}\left(C_L = 1 \mid A = a, W, S = s\right)\hat{\mathbb{P}}\left(A = a \mid W, S = s\right)\right\}^{-1}$ and using as offset logit of the predicted values from the initial model fit m'_{init}. Denote the fitted regression coefficient as $\hat{\epsilon}_H$, and the predicted values based on this model fit (which includes the offset) by $m'\left(W, \hat{\epsilon}_H\right)$.

Step 4: Let n denote the number of enrolled participants from subpopulation $S = s$. Estimate $P(Y = 1|A = a, S = s)$ by $\frac{1}{n}\sum_i m'\left(W_i, \hat{\epsilon}_H\right)$, where W_i is the baseline

variable for individual i, and the summation is over all enrolled participants from subpopulation $S = s$. This completes the TMLE algorithm.

Above, all logistic regression model fits use the R function glm, which implements maximum likelihood estimation. In step 3, even though H may not take integer values, it always takes values in the interval $[0, 1]$, and the quasibinomial glm family in R can be used to obtain a model fit.

Following the above algorithm of computing TMLE estimator, the subpopulation s average treatment effect δ_s is estimated by applying the above estimation algorithm to obtain estimates of $P(Y = 1|A = a, S = s)$ for each $a \in \{1, 0\}$ and then taking the difference between the former and the latter. The estimator of the average treatment effect for a composite population $\tilde{\mathcal{S}}_j$ is obtained analogously by replacing the condition $S = s$ by $S \in \tilde{\mathcal{S}}_j$ above, and using all participants in the composite population $\tilde{\mathcal{S}}_j$.

In Section (4.9.1), we referred to a version of the above TMLE estimator that uses the known quantities $P(A|W, S) = 1/2$ and the known fractions of total participants $i \in \{1, \ldots, N\}$ who have $C_{i,k}^{(t)} = 1$, i.e., the fraction $N_{s,k}^{(t)}/N$, rather than fitting models for these. This involves replacing each occurrence of $\hat{\mathbb{P}}(A|W, S)$, $\hat{\mathbb{P}}(C_L|A, W, S)$, and $\hat{\mathbb{P}}(C_Y|L, C_L, A, W, S)$ by the corresponding known quantity in steps 2-4. The reason for using estimates of these known quantities is that, as shown by van der Laan and Gruber (2012), this either improves or leaves unchanged the asymptotic precision of the resulting estimator.

We assumed that the only cause of missing data was administrative censoring due to some participants not yet having their final outcomes observed. It is straightforward to incorporate additional right censoring due to loss to follow-up, under the assumption of censoring at random, i.e., the sequential randomization assumption in (van der Laan and Gruber, 2012) stating that censoring is conditionally independent of the potential outcomes given the past observed data. The TMLE that we used already includes working models for the censoring variables $C_{i,k}^{(t)}$ conditional on a participant's data collected before time t. The TMLE could be applied unchanged, or the working models could be expanded to include additional terms if deemed necessary to explain loss to follow-up. Consistency of this TMLE (and any locally efficient estimator for this problem) under the above assumption requires that these working models are correctly specified. As described, e.g., by Lu and Tsiatis (2011), the unadjusted estimator is generally not consistent under the above assumption if censoring is informative, i.e., dropout is associated with the final outcome for reasons that can be attributed to variables such as side-effects measured after randomization. Using an adjusted estimator not only can improve precision but also leads to consistency under weaker assumptions than the unadjusted estimator.

4.9.3 Covariance Matrix Σ

In our simulations, for computational feasibility, we precomputed an approximation to the covariance matrix Σ, which we treated as known. This computation was based on Monte Carlo simulation, where we first generated 50,000 simulated trials with no early stopping so that the maximum number of participants were enrolled; in each simulated trial, the estimator and corresponding Wald statistic at each stage for each population were computed; the empirical covariance matrix of the 50,000 sets of statistics $\{Z_{j,k}\}_{j\leq J,k\leq K}$ was used to approximate Σ. Next, the algorithm in Section 4.5 was used to generate the efficacy boundaries $u_{j,k}$.

In practice, the covariance matrix Σ can be approximated by the nonparametric bootstrap using the data collected up through the current stage. Just as in the error-spending approach for standard, group sequential designs, it is only necessary to know (or approximate) the covariance of the statistics up through the current stage, in order to determine the thresholds $u_{j,k}$ for that stage. For the analysis at the end of stage k, we resample with replacement from all participants who have completed the trial to construct B replicated data sets $X^{*1}, \ldots, X^{*B}$, as described below. For each replicated data set X^{*b}, the sequence of stagewise estimators $\hat{\Delta}_{j,1}, \ldots, \hat{\Delta}_{j,k}$ and corresponding Wald statistics $Z_{j,1}, \ldots, Z_{j,k}$ are computed for each population $j \leq J$. We then compute the empirical covariance matrix of the set of Wald statistics over the B replicated data sets, which is a consistent estimator $\hat{\Sigma}^{(k)}$ of Σ restricted to the statistics at or before stage k. This estimator is used to construct the efficacy

boundaries $u_{j,k}$ for each $j \leq J$ based on (4.1).

For a given set of information levels $\mathcal{I}_{j,k}$, let $n_{s,k}$ denote the number of participants from subpopulation s who would have final outcomes observed during stage k, assuming no early stop for subpopulation s. Define $n_k = \sum_{s=1}^{m} n_{s,k}$, $n = \sum_{k=1}^{K} n_k$, and $\mathbf{n} = \{n_{s,k} : s = 1, \ldots, m; k = 1, \ldots, K\}$.

We next describe how each replicated data set X^{*b} is constructed at the stage k analysis for the unadjusted estimator. First, for each stage $k' \leq k$ and subpopulation s, $n_{s,k'}$ replicated participants are constructed with full data vector $(S = s, W, A, L^{(1)}, \ldots, L^{(T)}, Y)$, by drawing with replacement from the set of all subpopulation s participants who have complete data observed, i.e., all participants i enrolled in any stage at or before k such that $S_i = s$ and $C_{i,k}^{(T+1)} = 1$. Every subpopulation will have some such participants since each subpopulation s is enrolled in stage 1 and the first stage interim analysis occurs after $n_{s,1} > 0$ participants have final outcomes Y observed. We use the pooled set (over all stages $k' \leq k$) of subpopulation s participants with Y observed to draw replicated participants for each stage; this takes advantage of all outcome data accrued on subpopulation s. Implicit in this approach is the assumption of no change in the population distribution over time, which follows from the assumptions in Section 4.3; the assumption of no changes in the population distribution over time is typically needed even for standard, group sequential designs in order for hypothesis tests to be valid. This completes the description of how each X^{*b} is constructed; it is then used as

described above to estimate Σ restricted to the statistics at or before stage k. The construction of X^{*b} for the adjusted estimator is similar to that for the unadjusted estimator, except that pipeline participants are included.

Updates to the estimate of Σ based on data accrued at stage k may change the corresponding estimates of the probabilities on the left side of (4.1). This issue occurs for the error spending approach in standard, group sequential designs as well, as discussed for example by Scharfstein et al. (1997, Section 4.2). To address this issue, we recommend that a modified version of the algorithm in (4.1) be applied in which cumulative probabilities are used, i.e., at each interim analysis $k \leq K$, for each population j in $0, \ldots, J$ in turn, the efficacy boundary $u_{j,k}$ is defined as the solution to

$$P_{\hat{\Sigma}^k}\left(\text{for all } k' \leq k, j' \leq j : Z'_{j',k'} \leq u_{j',k'}\right) = 1 - \sum_{k' \leq k, j' \leq j} \alpha_{j,k}. \tag{4.15}$$

In the above equation, the values of $u_{j',k'}$ for stages $k' < k$ are those computed at stage k' based on the estimator $\hat{\Sigma^{k'}}$ that was available at that stage. The above algorithm is equivalent to using (4.1) if Σ is known, but may differ when using estimates of Σ updated after each stage. The above algorithm compensates, to a degree, for early estimates of Σ that led to setting efficacy boundaries $u_{j,k'}$ too low, by setting current stage boundaries higher (and vice versa).

In principle, it is possible that updated estimates of Σ could lead to no solu-

tion being possible for (4.15), which is similar to the case of nonmonotonicity of estimated information times in a standard, group sequential design using error spending approach (Scharfstein et al., 1997, Section 4.2). If this were to occur, we make a similar recommendation as Scharfstein et al. (1997) not to reject any new null hypotheses and to continue to the next stage if further enrollment is indicated under the enrollment rule r_k.

4.9.4 Construction of Data Generating Distribution in Section 4.6.2

In order to conduct a simulation study, we need a population to sample from. The 100 participants from the original MISTIE II data is a good candidate. However, although treatments were randomly assigned in the MISTIE Phase II trial, conditioning on the 100 participants, treatments are correlated to baseline covariates. In fact, for any baseline covariates, W, randomly chosen from the baseline covariates of the 100 participants, the corresponding treatment assignment is deterministic. To make treatment assignment independent of baseline covariates, we augment the data with a hypothetical twin for each participant. A participant has the same subpopulation membership and baseline covariates as his twin, but opposite treatment assignment.

When setting the short-term and final outcomes of the newly added twins,

there are two issues we should be specially wary of. First, to make the treatment effects realistic, the treatment effects in the augmented data should be comparable to those in the original MISTIE Phase II data. Secondly, since the semiparametric estimator leverages baseline covariates and short-term outcomes in estimating the mean final outcomes, the predictability of the final outcomes from the baseline covariates and short-term outcomes should be reasonable, in order not to exaggerate or underestimate the improvement of the semiparametric estimator.

Since at 30 and 90 days of treatment, patients with mRS ≤ 3 are rare, we let the short-term outcomes, $L^{(1)}$ and $L^{(2)}$, be the indicators of mRS no larger than 4 at 30 and 90 days of treatment, respectively. We generate the outcomes of the new added twins as follows. We first fit logistic regression models for $L^{(1)}$ on (W, A), for $L^{(2)}$ on $(L^{(1)}, W, A)$, and for Y on $(L^{(2)}, L^{(1)}, W, A)$, using the original MISTIE Phase II data. Then, preliminary short-term and final outcomes of the newly added twins are predicted based on these logistic regression models by truncating the success probabilities at 0.5. Next, to calibrate the treatment effects, for every newly added twin, we reset his final outcome to be $Y = A$ with probability 0.03. The resetting probability 0.03 is numerically solved using binary search, so that the treatment effect of the augmented data is enlarged to match the original data.

Next, we focus on calibrating the predictability of the generated final outcome. Since the semiparametric estimator also involves fitting generalized linear models to predict the final outcomes, care should be taken to avoid over fitting. We assume

that in the phase of data analysis, only part of $(W, L^{(1)}, L^{(2)})$ are available. In detail, we only observe $(W_1, W_2, L^{(1)})$, while $(W_3, L^{(2)})$ are missing. In the following, we set $W = (W_1, W_2)$ and $L = L^{(1)}$ for notational convenience. A good benchmark on the extent to which the semiparametric estimator may leverage (W, L) in predicting the mean final outcomes are the relative efficiencies against the unadjusted estimator under the original MISTIE Phase II data. To eliminate the correlation between A and (W, L), we reset A by a Bernoulli random variable, $B(0.5)$, exogenously in the original MISTIE Phase II data, and then apply the semiparametric estimator and the unadjusted estimator in five stages, assuming 1,000 overall sample size and equal increments between stages. The relative efficiencies are obtained by taking the ratio of the variances in estimating treatment effects. Then, under the augmented data, let $p_a = P(Y = 1|Y \sim Q_a)$ be the baseline success rate of treatment arm a. For every newly added twin, with probability 0.164, we reset his final outcome by a Bernoulli random variable $B(p_a)$. We note that this resetting step doesn't change the baseline success rates in each treatment arm, and thus doesn't change the treatment effect. However, it weakens the dependence between Y and (W, L). The resetting probability 0.164 is chosen so that, when treatment label A is reset by a Bernoulli random variable, $B(0.5)$, the relative efficiencies of the adjusted estimator against the unadjusted estimator using the augmented data mimic those using the original MISTIE Phase II data.

Ideally, we would assign each participant to two subpopulation groups based

on their IVH measure. However, this IVH measure is not available in the MISTIE II data set. Hence, for simulation purposes we randomly assign subpopulation information for each participant in our augmented data. When simulating trials under Scenario (b), we reassign the treatment A by Bernoulli(0.5) random variable for subpopulation 2 so that it has no treatment effect; when simulating trials under Scenario (c), we reassign the treatment A by Bernoulli(0.5) random variable for both the subpopulations. Note that since the subpopulation is for simulation purposes - to distinguish Scenario (a) and (b), and to demonstrate the advantage of adaptive enrichment design, its being merely a label suffices these ends. However, the baseline variable distribution is the same between two subpopulations - whether this is practically desired depends on the nature of a real trial.

The above data generating mechanism provides a hypothetical overall population of 200 participants, of which 100 come from the original MISTIE Phase II data and consist of fully measured $D = (W, A, L, Y)$, and the rest are hypothetical twins consisting of fully measured (W, A) but random (L, Y). At stages k certain number participants of each subpopulation are sampled from the overall population, according to Table 4.1.

Chapter 5

Phase III Studies: Lasso Estimation of Hierarchical Interactions for Analyzing Heterogeneous Treatment Effect

SUMMARY.[1]

Individuals differ in how they respond to a given treatment, and characterizing such variability in treatment response is an imperative aim in personalized medicine. We propose a general method to model treatment response heterogeneity, through identification of treatment-covariate interactions honoring different

[1]This Chapter 5 is adapted from the working paper "**Yu Du**, Ravi Varadhan. *Lasso Estimation of Hierarchical Interactions for Analyzing Heterogeneous Treatment Effect*."

hierarchy conditions. We construct a single-step l_1 norm penalty procedure that maintains the hierarchical structure of interactions in a sense that the treatment-covariate interaction term is included in the model only when either of the covariate or the covariate and the treatment both have non-zero main effects. We explore several parameterization schemes with different constraints added to Lasso that enforce the hierarchical interaction restriction. We solve the resulting constrained optimization problem using a spectral projected gradient method. We compare our methods to the unstructured Lasso using simulation studies covering a variety of scenarios for treatment-covariate interactions. The simulations show that our methods yield more parsimonious models and outperform unstructured Lasso in terms of prediction performance, and in terms of the ability to correctly identify non-zero treatment covariate interactions. The superior performance of our methods are also corroborated by an application on a large randomized clinical trial data investigating a drug for treating congestive heart failure (N=2,569). Our methods can be applied to continuous, binary and time to event outcome, providing a well-suited approach with sufficient flexibility in terms of parametrization for doing secondary analysis in clinical trials to analyze heterogeneity treatment effect.

5.1 Background

Individuals differ in how they respond to a given treatment, and characterizing such variability in treatment response is an important aim when practicing personalized medicine. This heterogeneity in treatment response is well recognized in clinical practice. Hence any summary from a clinical trial, such as the overall treatment effect, is not directly relevant for treating individual patients (Rothwell, 1995; Bailey, 1994; Kent and Hayward, 2007). A new treatment might only benefit a sub-population of the patients with certain characteristics, while it shows no benefit to others. Accurate evaluations of this heterogeneity attributable to the variation in baseline patient characteristics provides many potential benefits in terms of facilitating the decision-making in appropriately targeting existing therapies to the individuals. Thus, the investigation of patient heterogeneity in treatment response becomes an imperative, which answers the question of what characteristics of the patients are associated with a benefit from the studied treatment. One of the current approaches for assessing treatment response heterogeneity is to find predictive factors for the treatment, which are covariates predictive of treatment response. In Epidemiology such variables are called effect modifiers, namely, the treatment effect will vary across individuals for different values of these variables. Predictive factors are often derived from prognostic factors, the variables imposing an impact on the outcome in the absence of the treatment. Often, there are numerous prognostic factors, and to find which ones are predictive of treatment

response remains an important task.

There is a robust literature addressing the estimation of heterogeneity of treatment effect including e.g., Jiang et al. (2007), Zhou et al. (2008), Barker et al. (2009), Freidlin et al. (2010), Lee et al. (2010), Kim et al. (2011), Cai et al. (2011), Lai et al. (2014), Xu et al. (2014), Ohwada and Morita (2016), Spencer et al. (2016), Henderson et al. (2016). A common strategy for studying the treatment response heterogeneity in clinical trial is subgroup analysis (Rothwell, 2005), which explores how treatment effect varies across subgroups defined by one variable at a time. Therefore, the subgroup analysis ignores the joint effect of the covariates on treatment effect. Hence, it may fail to identify significant treatment covariate interactions, especially when the number of variables is fairly large, which is often the case in clinical trial studies. Another common approach would be to pre-specify all the prognostic covariates and fit a model with all treatment-covariate interactions. This is termed as unstructured interaction model (Kovalchik et al., 2013). However, this approach has two important limitations. The variables need to be pre-specified, and the model may include interactions that are not actually present.

Kovalchik et al. (2013) proposed a parsimonious approach, extending the work of Follmann and Proschan (1999), to use proportional interactions model to investigate treatment response heterogeneity in a randomized controlled clinical trial. This overcomes the limitation of subgroup analysis by jointly considering the effect modification of various variables, however, it still has difficulty in dealing

with fairly large number of candidate effect modifiers. Another limitation is that of model misspecification when the underlying treatment-covariate interactions are not proportional to the main effects.

We propose a general method in this chapter that overcomes the limitations of Kovalchik et al. (2013), to assess heterogeneity of treatment response in clinical trials setting. Our work differs in two aspects: (1) We relax the "proportional" constraint, allowing more flexibility in estimating treatment covariate interactions; (2) the methods are able to automatically screen a large number of potential effect modifiers, with desirable properties of low false-positives and false-negatives to correctly capture the significant interactions.

Our work is also inspired by the work of Bien et al. (2013), which modifies Lasso (Tibshirani, 1996) to estimate a sparse interaction model, considering the complete list of two-way interactions. One important constraint Bien et al. (2013) has employed is the interaction hierarchy restriction that an interaction term can only be included in the model when one or both of the associated variables are retained in the model. Our proposed methods make use of this principle, which is considered very practical, as Cox (1984) once stressed that "Large component main effects are more likely to lead to appreciable interactions than small components. Also, the interactions corresponding to larger main effects may be in some sense of more practical importance."

Our work differs from Bien et al. (2013) in two ways. First, we focus on treatment-

covariate interactions in the context of randomized clinical trials, whereas Bien et al. (2013) examines all two-way interactions in a context different than evaluating treatment effect heterogeneity. Second, we propose four different ways to incorporate hierarchical constraints, two of which enforce the hierarchy constraints in a direct manner. Thus, our work provides a new and potentially useful framework for identifying predictive variables for heterogeneous treatment response.

Given a large set of p covariates from a randomized clinical trial, we select and estimate a subset of candidate effect modifiers that are predictive of treatment response. Our modeling approach is fully parametric and provides a clear interpretation of how individual baseline characteristics affect treatment response. The primary outcome we consider in this chapter is time-to-event, although our proposed methods are easily adapted to other types of endpoints, e.g., continuous or binary. In Section 5.2, we introduce the notation, interaction hierarchy restriction, and the parametric modeling assumption. We construct a single-step l_1 norm penalty procedure that maintains the hierarchical structure of interactions. We study four parameterization schemes with different hierarchy constraints added to Lasso. The formulated constrained optimization problems are solved by using the spectral projected gradient method. We examine the performance of our proposed methods with simulation studies in Section 5.3 in the presence of interaction hierarchy or non-hierarchical interactions and compare with Lasso (unstructured). In Section 5.4, we apply our method to the Studies of Left Ventricular Dysfunc-

tion Treatment (SOLVD-T) trial, a two-arm placebo-controlled randomized clinical trial investigating the efficacy of enalapril, the angiotensin-converting-enzyme inhibitor, to reduce the hazard of death or hospitalization among patients with chronic heart failure (The-SOLVD-Investigators, 1991). We provide a discussion in Section 5.5 of the proposed methods, and discuss future extensions of this work.

5.2 Method

5.2.1 Notations and Assumptions

In this chapter, we target the time-to-event (TTE) outcome, where we use T to denote the event time, and C, the censoring time. The non-informative censoring is assumed throughout. The observed outcome is represented by a vector (X, Δ), $X = \min(T, C)$, and $\Delta = \mathrm{I}(T \leq C)$, where $\mathrm{I}(T \leq C)$ is the indicator variable taking value 1 if $T \leq C$. We are dealing with the context of a parallel-arm clinical trial, so let A be the treatment indicator, where $A = 1$ means assignment to the treatment arm while $A = 0$ means assignment to the control arm. Let Z be the vector of p candidate effect modifiers, such that $Z = (Z_1, Z_2, \ldots, Z_p)'$. Thus, the complete observed data for subject i is denoted by the vector $D_i = (X_i, \Delta_i, A_i, Z_i)$, $i = 1, 2, \ldots, n$, assuming that there are n observations in total. Aiming to capture the treatment covariate interactions, we assume the following multivariate Cox

proportional hazards model (Cox, 1972) to relate the TTE outcome to the subject's treatment assignment, p candidate prognostic factors as well as p candidate interaction terms:

$$\lambda(t|A_i, Z_i) = \lambda_0(t) \exp\left(\beta_A A_i + \beta_Z{}' Z_i + \gamma' A_i Z_i\right), \tag{5.1}$$

where $\lambda_0(t)$ represents baseline hazard at time t for any subject i drawn from the overall population, $\lambda(t|A_i, Z_i)$ gives the hazard function at time t conditioned on the subject i treatment assignment and his / her candidate prognostic factors (candidate effect modifiers). In this Cox model (5.1), β_A acts as the main effect of treatment, interpreted as the log hazard ratio comparing treatment to control, while $\beta_Z = (\beta_{Z_1}, \beta_{Z_2}, \ldots, \beta_{Z_p})'$ is a p element vector denoting the prognostic effect of Z, and $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_p)'$ is also a p element vector showing the treatment covariate interaction effect. Any non-zero $\gamma_j, j = 1, 2, \ldots, p$ in the vector γ identifies an effect modification between the treatment A and the corresponding prognostic factor Z_j in a sense that treatment effect depends on the factor Z_j, causing the heterogeneity in treatment effect across subjects in the population. The advantage of assuming a fully parametric model incorporating pairwise interactions between treatment and covariates is that we can provide a clear interpretation of how individual baseline characteristics affect treatment response. For example, for the prognostic factor Z_j, $\gamma_j > 0$ indicates a heterogeneity in a direction that the treatment becomes less efficacious for higher values of Z_j while $\gamma_j < 0$ implies a stronger treatment response when Z_j increases. Let us use θ to represent the vector of all the parameters, such

that $\theta = (\beta_A, \beta_Z{}', \gamma')'$. Here we use the TTE outcome as an illustration of the methods, however, the optimization problems we shall discuss can be easily adapted to a broader class of linear models, in addition to the Cox model (5.1), including the generalized linear models, where the primary outcome can either be continuous or binary.

5.2.2 Parameterization Schemes and Optimization Problems

Interaction hierarchy restriction states that an interaction term is only allowed into the model if one or both of the corresponding variables are marginally important, namely, the variable(s) has non-zero main effect in the model. Various names have been called for such restriction by Chipman (1996), Nelder (1977) and Peixoto (1987) among others, for example, “heredity" and “marginality". We adapt the definition to the clinical setting in the context of a parallel-arm clinical trial, with the goal to capture treatment covariate interactions as precisely as possible. Therefore, we define a hierarchical structure for treatment covariate interaction such that for $j = 1, 2, \ldots, p$

$$\gamma_j \neq 0 \Longrightarrow \text{at least } \beta_{Z_j} \neq 0. \tag{5.2}$$

We make direct use of this hierarchy in the parameterization schemes that will

follow. Traditionally, the parameter vector θ is estimated by minimizing over θ without any constraints the negative log partial likelihood (namely maximizing the partial likelihood), as defined by $l(\theta)$ such that

$$\begin{aligned} l(\theta) &= -\sum_{i:\Delta=1} \log \frac{\lambda(X_i|A_i, Z_i)}{\sum_{j:X_j \geq X_i} \lambda(X_i|A_j, Z_j)} \\ &= -\sum_{i:\Delta=1} \left(\beta_A A_i + \beta_Z{}' Z_i + \gamma' A_i Z_i - \log \sum_{j:X_j \geq X_i} \exp\left(\beta_A A_j + \beta_Z{}' Z_j + \gamma' A_j Z_j\right) \right). \end{aligned}$$

It is often the case that there are only a handful of estimates out of many that have non-zero effects in the model, corresponding to a sparse structure. To address this, Lasso is a widely applied technique proposed by Tibshirani (1996) that employs model selection and estimation at the same time. This is accomplished by imposing an l_1 norm penalty on the parameter vector and solving a constrained convex optimization problem:

$$\begin{aligned} \underset{\theta}{Minimize} \quad & l(\theta) \\ \text{subject to} \quad & \|\theta\|_1 \leq \lambda, \end{aligned} \tag{5.3}$$

where λ acts as the penalty parameter, balancing the tradeoff between fitting to the training data and the enforcement of a sparsity coefficients structure. The less the value of λ is, the more sparse the estimates of the parameter vector θ will be. We build interaction hierarchy into the optimization problem (5.3) by proposing

four parameterization schemes as constraints to extend Lasso.

- Parameterization scheme 1

$$\begin{aligned} \underset{\theta}{Minimize} \quad & l(\theta) \\ \text{subject to} \quad & \gamma_j = \beta_A * \beta_{Z_j} * \zeta_j, \text{ for } j = 1, 2, \ldots, p, \\ & \|\beta_A\|_1 + \|\beta_Z\|_1 + \|\zeta\|_1 \leq \lambda, \end{aligned} \tag{5.4}$$

  where $\zeta = (\zeta_1, \zeta_2, \ldots, \zeta_p)'$ is a p element vector. The added constraint $\gamma_j = \beta_A * \beta_{Z_j} * \zeta_j$ enforces "strong hierarchy" as defined by Bien et al. (2013), since it is straightforward to note that if for any particular j, $\hat{\gamma}_j \neq 0$, then both the treatment effect and the prognostic effect of Z_j have significant (non-zero) estimates, $\hat{\beta}_A \neq 0$ and $\hat{\beta}_{Z_j} \neq 0$.

- Parameterization scheme 2

$$\begin{aligned} \underset{\theta}{Minimize} \quad & l(\theta) \\ \text{subject to} \quad & \gamma_j = \beta_{Z_j} * \zeta_j, \text{ for } j = 1, 2, \ldots, p, \\ & \|\beta_A\|_1 + \|\beta_Z\|_1 + \|\zeta\|_1 \leq \lambda. \end{aligned} \tag{5.5}$$

This scheme serves as a weaker version of scheme 1 since it guarantees only that $\hat{\beta}_{Z_j} \neq 0$ if $\hat{\gamma}_j \neq 0$, lifting the imposing of β_A on the parameterization of γ. Our intention is that a relaxation of the constraint on γ would bring more robustness to the estimation of the parameters, which we shall explore in the simulation study. Note that Parameterization scheme 1 and 2 still maintain the convexity of the constraints inherited from Lasso, though they both increase the non-linearity of the optimization problem. Given the non-linear nature of the partial likelihood function and that there is no closed-from solution for Lasso, the increased non-linearity is computationally tractable and not undesirable to solve.

- Parameterization scheme 3

$$\begin{aligned} \underset{\theta}{Minimize} \quad & l(\theta) \\ \text{subject to} \quad & |\gamma_j| \leq \min\left(|\beta_A|, |\beta_{Z_j}|\right), \text{ for } j = 1, 2, \ldots, p, \\ & \|\theta\|_1 \leq \lambda. \end{aligned} \tag{5.6}$$

Instead of parameterizing the interaction coefficients γ, we impose the inequality on γ to enforce the hierarchical interaction restriction in scheme 3. Note that if $|\hat{\gamma}_j| > 0$, so are $|\hat{\beta}_A|$ and $|\hat{\beta}_{Z_j}|$.

- Parameterization scheme 4

$$\begin{aligned} \underset{\theta}{Minimize} \quad & l(\theta) \\ \text{subject to} \quad & |\gamma_j| \leq |\beta_{Z_j}|, \text{ for } j = 1, 2, \ldots, p, \\ & \|\gamma\|_1 \leq |\beta_A|, \\ & \|\theta\|_1 \leq \lambda. \end{aligned} \tag{5.7}$$

Scheme 4 is adapted from the added constraint in Bien et al. (2013) to fit the clinical setting, dealing only with the treatment covariate interactions. This imposes a stronger penalization on the interaction parameters γ so that a more sparse structure of interaction coefficient estimates shall be produced than scheme 3. Notice that the condition in scheme 3, $|\gamma_j| \leq |\beta_A|$, for any j, $j = 1, 2, \ldots, p$, is replaced in scheme 4 by the constraint on the norm of the entire vector γ, such that $\|\gamma\|_1 = \sum_j |\gamma_j| \leq |\beta_A|$, resulting in a bigger penalty.

All the above four parameterization schemes we propose are modifications of Lasso (Tibshirani, 1996) that guarantee to produce models satisfying interaction hierarchy restriction.

5.2.3 Algorithm to Solve the Optimization Problems

Despite the enforcement of the interaction hierarchy restriction, the optimization problems with scheme 3 and scheme 4 listed in Section 5.2.2 are not convex, thus undesirable to solve. We thus implement simple convex relaxation of the problems by substituting each of the variable with two of the components such that for each $j = 1, 2, \ldots, p$,

$$\gamma_j = \gamma_j^+ - \gamma_j^-, \tag{5.8}$$

$$\beta_{Z_j} = \beta_{Z_j}^+ - \beta_{Z_j}^-, \tag{5.9}$$

$$\beta_A = \beta_A^+ - \beta_A^-, \tag{5.10}$$

$$|\gamma_j| = \gamma_j^+ + \gamma_j^-, \tag{5.11}$$

$$|\beta_{Z_j}| = \beta_{Z_j}^+ + \beta_{Z_j}^-, \tag{5.12}$$

$$|\beta_A| = \beta_A^+ + \beta_A^-, \tag{5.13}$$

where $\gamma_j^+ \geq 0, \gamma_j^- \geq 0, \beta_{Z_j}^+ \geq 0, \beta_{Z_j}^- \geq 0, \beta_A^+ \geq 0, \beta_A^- \geq 0$. Note that γ_j^+, $\beta_{Z_j}^+$, β_A^+, γ_j^-, $\beta_{Z_j}^-$, β_A^- are not defined to be the positive parts and negative parts of the variables since we do not add the constraints $\gamma_j^+\gamma_j^- = 0$, $\beta_{Z_j}^+\beta_{Z_j}^- = 0$, and $\beta_A^+\beta_A^- = 0$, which makes the problem convex and tractable to solve. Another consequence out of this is to make the constraints less restrictive because for example, it is possible that the solution to such optimization problems returns both $\hat{\beta}_{Z_j}^+ > 0$ and $\hat{\beta}_{Z_j}^- > 0$ for the estimate of β_{Z_j}, therefore, the condition like $\gamma_j^+ + \gamma_j^- \leq \beta_{Z_j}^+ + \beta_{Z_j}^-$ might

have a larger bound. This convex relaxation still guarantees interaction hierarchy, as proved in Bien et al. (2013). For scheme 1 and scheme 2, it is an advantage that the added constraints do not alter the convexity of the problem, though a potential disadvantage is that it increases the nonlinearity of the problem, however, this is not a major issue computationally since the objective function, the partial log-likelihood, is already nonlinear.

We applied the spectral projected gradient (SPG) method, proposed by Birgin et al. (2000), to solve the optimization problems in Section 5.2.2. As its name suggests, the SPG method incorporates the spectral gradient scheme (Raydan, 1997) to greatly improve the effectiveness of the gradient projection method (Bertsekas, 1976 and references therein). As pointed out in Birgin et al. (2014), the SPG method has been applied in wide areas of statistics, becoming an ideal tool for large-scale convex constrained optimization problems. The general SPG algorithm used in this chapter is listed in Table 5.1, where θ denotes the vector of all the parameters in the model as defined in Section 3.2.2, however, for scheme 1 and 2 in Section 5.2.2, $\theta = (\beta_A, \beta_Z{}', \zeta')'$, due to the parameterization of γ.

The computation of the step length α_k and the spectral step length λ_k are given in details in Birgin et al. (2014). An implementation of the SPG method is readily available in R (R Core Team, 2016), within the **BB** package written by Varadhan and Gilbert (2009). The function **spg**() in the **BB** package provides 3 different options for spectral step lengths: 1) the step length used in Birgin et al. (2000); 2) the

Input:
$l(\theta)$, the negative log partial likelihood of survival data;
$\nabla l(\theta)$, the gradient function of $l(\theta)$;
λ_0, the initial value of the spectral step length;
$\mathbb{P}_\Omega()$, the projection function into the convex set Ω of the constraints;
$\theta_0 \in \Omega$, the initial value of θ;
ϵ, the value of tolerance.
Goal:
find the minimizer $\hat{\theta}$ of $l(\theta)$ subject to $\hat{\theta} \in \Omega$.
Algorithm:
at the k^{th} iteration, $k \geq 1$,
while not convergent, e.g., $\left\|\mathbb{P}_\Omega\left(\theta_k - \nabla l(\theta_k)\right) - \theta_k\right\|_\infty > \epsilon$,
do:
compute the search direction $d_k = \mathbb{P}_\Omega\left(\theta_k - \lambda_k \nabla l(\theta_k)\right) - \theta_k$,
compute the step length α_k,
compute $\theta_{k+1} = \theta_k + \alpha_k d_k$,
compute spectral step length λ_{k+1}.

Table 5.1: Algorithm for SPG method

step length proposed in Barzilai and Borwein (1988); 3) the step length proposed by Varadhan and Roland (2008). We used the function **spg**() with the third choice of spectral step length, as recommended by Varadhan and Gilbert (2009), to carry out the method of SPG, solving the convex constrained optimization problems in Section 5.2.2. Using the function **spg**() requires us to provide the computation for $l(\theta)$, the negative log partial likelihood of survival data; $\nabla l(\theta)$, the gradient function of $l(\theta)$; and $\mathbb{P}_\Omega()$, the function to project any arbitrary point into the feasible convex set Ω of the constraints. These are straightforward to derive and code for the four parameterization schemes and the optimization problems listed in Section 5.2.2. Once all these arguments are provided, the function returns the minimizer $\hat{\theta}$ of the function $l(\theta)$, and any non-zero $\hat{\gamma}_j$ reveals a significant interaction between the

treatment and the prognostic factor Z_j. A complete list of codes solving the proposed optimization problems can be found on the author's Github account.

There are also other methods to solve these convex constrained optimization problems, for example, a sequential quadratic programming algorithm written by Kraft (1988), among others. We explored that sequential quadratic programming method in the simulation study and it produces very similar results and performance to SPG method. The focus of the chapter is the proposal of the four parameterization schemes that extend Lasso and build the interaction hierarchy restriction into the model for the clinical setting, dealing with the identification of the treatment covariate interactions out of many candidate effect modifiers. Though this chapter does not aim to find the most efficient method for solving these convex constrained optimization problems, it remains an open area for future work.

5.3 Simulation

5.3.1 Simulation Setup

In this section, we conduct several simulations to study the leverage of the interaction hierarchy restriction in the context of four parameterizations listed in Section 5.2.2. A hundred clinical trials are simulated such that each trial assigns $n = 500$ patients to treatment and control arm with 1:1 randomization ratio. Each

trial comes with $p = 50$ potential prognostic factors, thus, 50 candidate treatment covariate interactions are expected. We let 25 prognostic factors out of 50 have significant impact on the outcome (i.e., with non-zero coefficients). We set 5 treatment covariate interaction coefficients to be non-zero. The primary outcome is time-to-event (TTE) survival endpoint, generated by exponential distribution assuming the proportional hazard model (5.1), whereas the censoring time is generated from an independent exponential distribution. The simulation study aims to evaluate the performance of these methods for identifying non-zero treatment covariate interactions in the presence of hierarchical interaction and non such interaction hierarchy. Therefore, we mainly consider two scenarios for the underlying data generating distribution:

(A) Hierarchical interaction is enforced,

$$\gamma_j \neq 0 \quad \Rightarrow \quad \beta_{Z_j} \neq 0 \quad \text{and} \quad \beta_A \neq 0,$$

(B) No interaction hierarchy restriction.

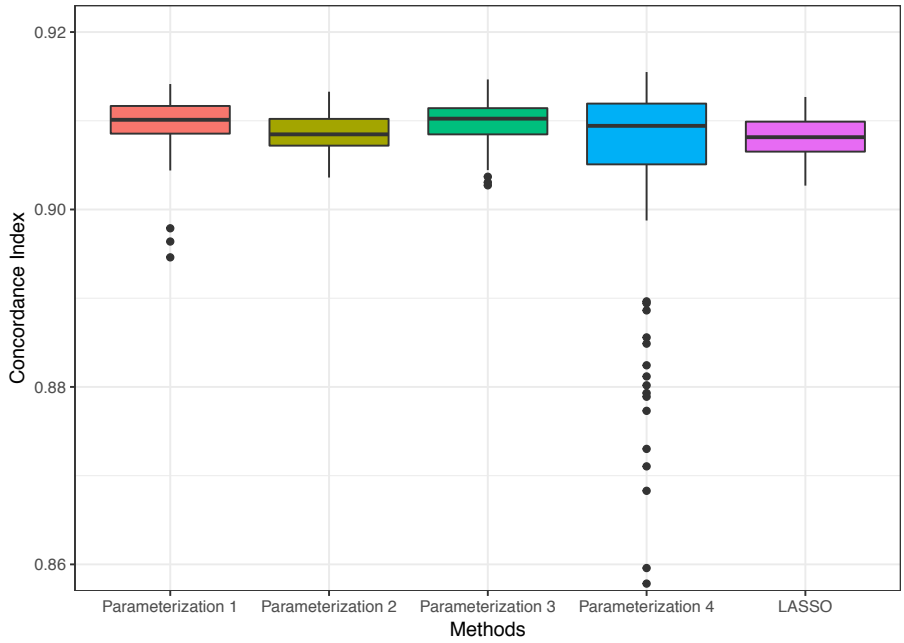
5.3.2 Simulation Evaluation and Results

Not only we assess the performance when the fundamental assumption about the interaction hierarchy is satisfied, but we also consider the case where it is misspecified to evaluate the model's sensitivity to this assumption. Lasso, unmod-

ified, serves as a basis of comparison since this is a simple and straightforward approach most commonly used by statisticians for parameter regularization and identification of a sparse structure of the coefficients. The evaluation of performance is twofold: (1) prediction performance and (2) the ability of models to correctly recover the non-zero treatment covariate interactions. In the assessment of prediction performance, we choose to use the concordance index, as shown by Harrell et al. (1996) and Pencina and D'Agostino (2004). The concordance index is a common metric to evaluate the risk prediction for TTE outcome, and if a pair of comparable subjects are drawn randomly from the population, it represents the probability that the subject with higher predicted risk would experience the event before the other one. For each of the methods, the penalization parameter λ should be set so that the model can be estimated. We apply 10 fold cross validation on each 500 subjects simulated trial in search of the λ that corresponds to the highest concordance index using the out-of-sample risk predictions for these 500 subjects. Accompanying each simulated trial is an invisible trial where we generate another 10000 subjects data under the same mechanism, serving as the validation set and we apply each determined model to this validation set to compute a concordance index as the assessment of prediction performance for each method. We repeat this process 100 times and summarize the results in Figure 5.1 for scenario (A) and (B).

Figure 5.1a displays the performance in risk prediction via the boxplots of concordance index for each of the methods in scenario (A) where the hierarchical in-

(a) The boxplots of concordance index in scenario (A)



(b) The boxplots of concordance index in scenario (B)

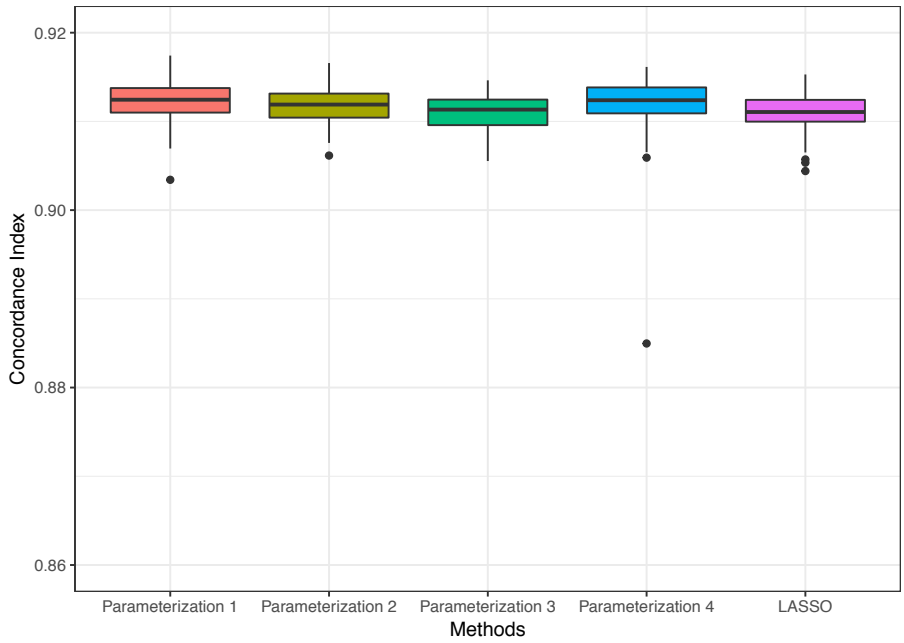


Figure 5.1: The comparison in risk prediction using concordance index.

teraction restriction is enforced in the data generating distribution while that for scenario (B) is represented in Figure 5.1b. We can clearly see that in both scenarios all these methods have comparable risk prediction accuracy, and on average (median), all the four parameterizations listed in Section 5.2.2 that incorporate interaction hierarchy have greater concordance index than Lasso, though Parameterization scheme 4 (5.7) is highly left skewed in scenario (A). This gives us the reassurance for the proposed methods since even in scenario (B) where the truth has no interaction hierarchy, our proposed methods can still rival Lasso in terms of risk prediction.

The main advantage of the proposed methods lies in the ability to correctly recover the significant, namely non-zero, treatment covariate interactions. To demonstrate this ability, we assess the sensitivity and the specificity of each model for the identification of treatment covariate interaction terms, and in addition, we also engineered a global metric of performance, called Global Interaction Recovery Cost (GIRC), to provide a single metric combining the sensitivity and specificity. Once the model is determined using cross validation, the parameter estimates are acquired and it is straightforward to compute the sensitivity and the specificity regarding the interaction terms. Figure 5.2 shows the boxplots of the sensitivity (Figure 5.2a) and the specificity (Figure 5.2b) for each method regarding the identification of treatment covariate interactions in scenario (A). As shown in Figure 5.2a, the sensitivity of recovering non-zero interactions for Parameterization scheme 3

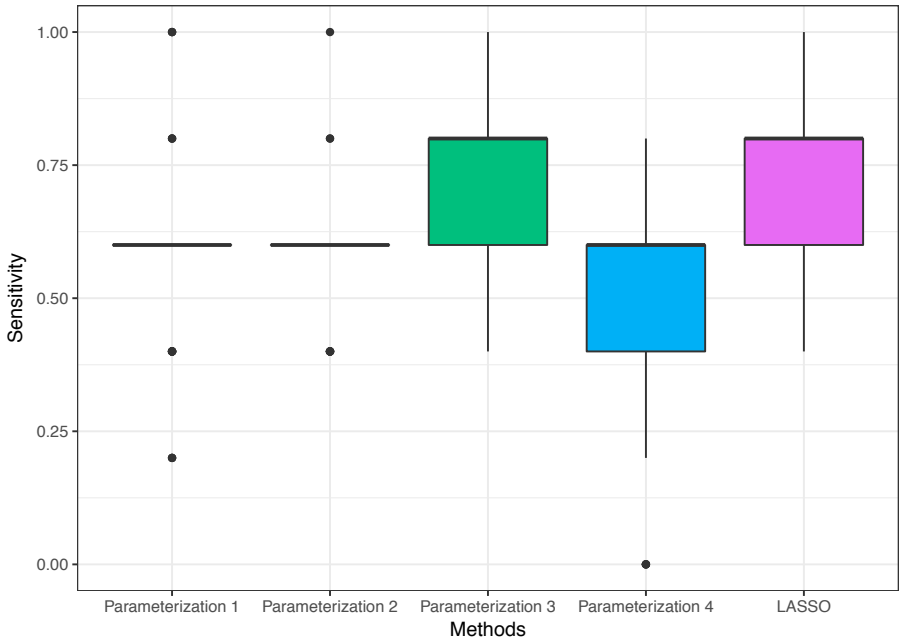
(5.6) is comparable to Lasso while that of the other parameterizations (5.4), (5.5) and (5.7) lies below Lasso. However, remember that the simulation sets up only 5 non-zero treatment covariate interactions, and the difference between the proposed methods and Lasso in averaged sensitivity amounts to only one term difference. That means on average, Lasso can recover 4 out of 5 non-zero interactions, the same as Parameterization scheme 3 (5.6) while the other parameterizations are also able to identify 3 out of 5. When it comes to the specificity of the interaction terms as displayed in Figure 5.2b, that is where the advantage of these methods is appreciated, especially when we are faced with numerous unknowingly false effect modifiers in practice, as in this simulation. Apart from Parameterization scheme 3 (5.6) which shows only a little improvement over Lasso, all others have at least around 15% greater specificity than Lasso, that amounts to $15\% * (50 - 5) \approx 7$ more terms correctly labeled as no effect modification. Thus, the treatment covariate interactions identified by the proposed methods are more likely to be the true effect modifiers than those given by Lasso. This is what happens when the hierarchical interaction restriction is satisfied in truth. Figure 5.3 exhibits the corresponding performance of each method in the sensitivity (Figure 5.3a) and the specificity (Figure 5.3b) for scenario (B) where the truth has no such interaction hierarchy in place. Each method has about the same averaged sensitivity, while for specificity there is still some little improvement for Parameterizations (5.4), (5.5) and (5.6) over Lasso on average, except a huge difference, 20%, favoring Parameterization

scheme 4 (5.7).

5.3.3 Global Metric of Performance

We propose a global metric of performance, Global Interaction Recovery Cost (GIRC), summarizing the sensitivity and specificity of the models for the identification of treatment covariate interaction terms. The sensitivity encodes the avoiding of false negative interaction terms while the specificity does the same for false positive interaction terms. Let $\mathcal{N}_{\text{FP}}$ and $\mathcal{N}_{\text{FN}}$ denote the number of false positive and false negative treatment covariate interaction terms, respectively. We use $\mathcal{N}$ to represent the total number of candidate interaction terms. Two types of error are associated with the false positive and false negative interaction terms, error of commission and error of omission. The error of commission is committed when the irrelevant interactions are included in the model (false positive) while the error of omission is committed when the relevant interactions are excluded from the model (false negative) . Let C_1 and C_2 be the cost associated with committing the error of commission per included irrelevant interaction term and committing the error of omission per excluded relevant interaction term, respectively. We assume that C_1 and C_2 are constants that may reflect monetary cost, time cost, resource cost, and etc., for the two types of error. The Global Interaction Recovery Cost

(a) The boxplots of sensitivity for treatment covariate interactions



(b) The boxplots of specificity for treatment covariate interactions

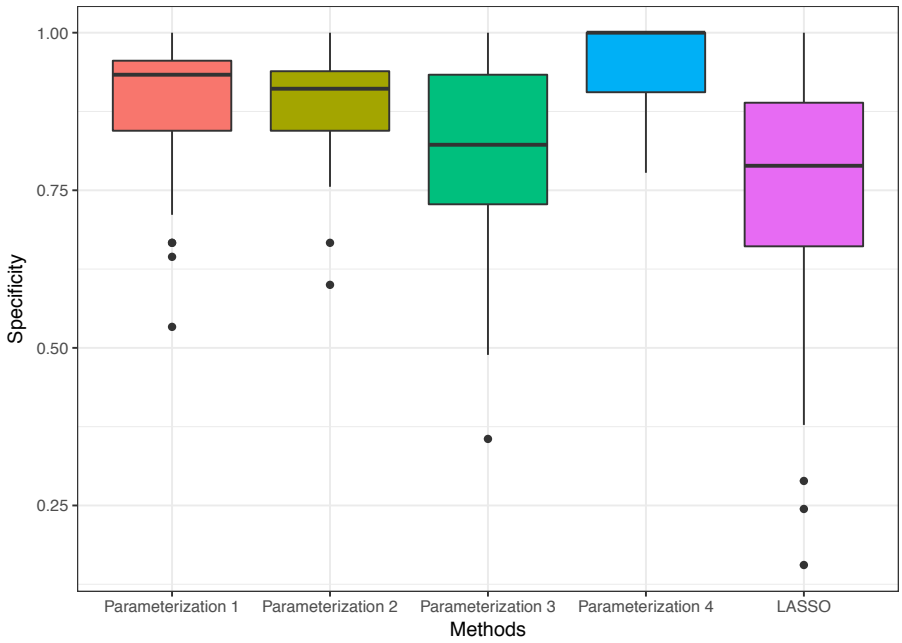
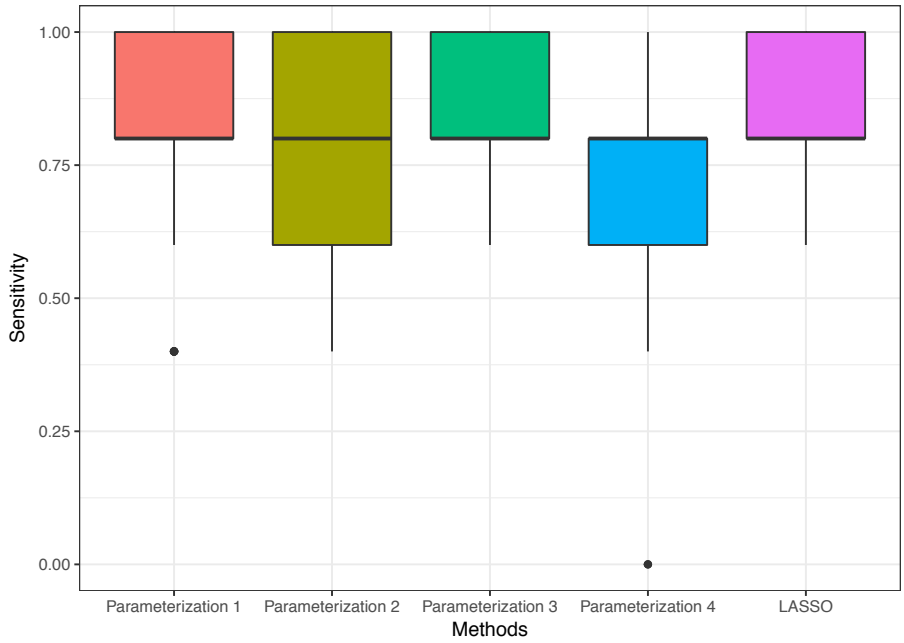


Figure 5.2: The ability to recover non-zero interactions in scenario (A).

(a) The boxplots of sensitivity for treatment covariate interactions



(b) The boxplots of specificity for treatment covariate interactions

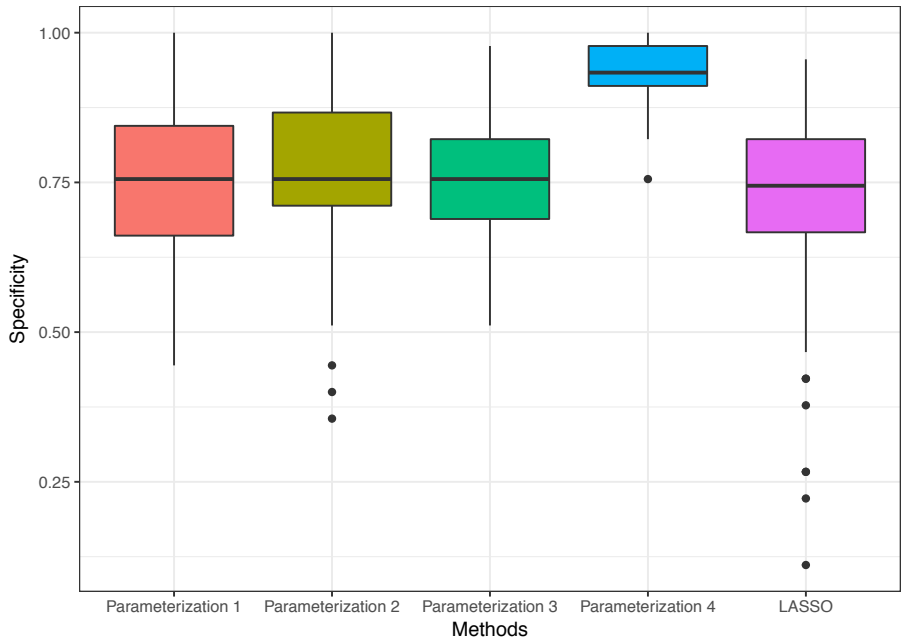


Figure 5.3: The ability to recover non-zero interactions in scenario (B).

(GIRC) is constructed such that

$$\text{GIRC} = 1/\mathcal{N}\left(C_1\mathcal{N}_{\text{FN}} + C_2\mathcal{N}_{\text{FP}}\right), \tag{5.14}$$

where C_1 and C_2 can be normalized such that $C_1 + C_2 = 1$. Therefore, GIRC is a global metric of performance, combining the sensitivity and the specificity and informing the combined cost of recovering non-zero interaction terms. The greater ability of capturing the significant treatment covariate interactions is warranted by the smaller value of GIRC. Assuming $C_1 = C_2 = 0.5$ where the two types of error incur the same amount of cost per term, we compute the average GIRC for each of the methods in scenario (A) and (B), which are summarized in Table 5.2.

Table 5.2: The average Global Interaction Recovery Cost (GIRC) of the identification of treatment covariate interactions for each method in scenario (A) and (B). P1, P2, P3 and P4 represents Parameterization scheme 1, 2, 3 and 4 respectively. $C_1 = 1/2, C_2 = 1/2$.

	P1	P2	P3	P4	Lasso
Scenario (A)	0.05	0.06	0.09	0.02	0.10
Scenario (B)	0.12	0.12	0.12	0.04	0.13

GIRC, as computed in this simulation study, demonstrates the superior performance of our proposed methods against Lasso, especially in truth where there is interaction hierarchy in place, and the ratio of GIRC comparing Lasso to our proposed methods ranges from 1.1 – 5. Remember we assume that two types of error incur equal cost, $C_1 = C_2 = 0.5$, however, in practice, the two types of cost are very likely to differ. Intuitively, committing error of commission should be more costly

in that the resulting false positive treatment covariate interactions shall give rise to unnecessary external validation trial, a waste of resources, and misguide the treatment recommendation that might bring harm to the targeted patients. Therefore, it is often the case that $C_1 \leq C_2$. As we see in Figure 5.2b and Figure 5.3b, the advantage of using our proposed methods are mainly channeled through the greater specificity compared to Lasso, i.e., the fewer false positive interaction terms. Table 5.3 shows the average GIRC when the cost of committing error of commission is twice that of committing error of omission, $C_2 = 2C_1$. The maximum ratio of GIRC comparing Lasso to our methods rises up from 5 to 13. Hence, we believe that in practice, our proposed methods are even more advantageous over Lasso.

Table 5.3: The average Global Interaction Recovery Cost (GIRC) of the identification of treatment covariate interactions for each method in scenario (A) and (B). P1, P2, P3 and P4 represents Parameterization scheme 1, 2, 3 and 4 respectively. $C_1 = 1/3, C_2 = 2/3$.

	P1	P2	P3	P4	Lasso
Scenario (A)	0.05	0.07	0.11	0.01	0.13
Scenario (B)	0.15	0.15	0.15	0.05	0.16

5.4 Data Application

5.4.1 Data Description

We evaluated our proposed methods on a completed, placebo-controlled, randomized trial (SOLVD-T) that tested the efficacy of an experimental drug, enalapril,

the angiotensin-converting-enzyme inhibitor, for treating chronic heart failure patients (The-SOLVD-Investigators, 1991). 1284 patients were assigned randomly to the control arm while 1285 to the treatment arm. The primary outcome of the trial is a TTE endpoint, the time to hospitalization or death. We assumed non-informative censoring, and around 47% of the outcome were censored. There are 23 candidate effect modifiers, including baseline age, gender, New York Heart Association (NYHA) function status, sodium level, creatinine level, etc. The goal of our proposed methods for this study is to select a subset of these candidates that are predicted to have non-zero interactions with treatment, and to provide the estimates. Missing values in the dataset are approached by imputation, where Missing at Random (MAR) are assumed. For example, if Z_1 has missing values, we regress Z_1 on the observed values of all other covariates $Z_2, \ldots, Z_p$ so that the missing values are filled by the predictive value from the regression. This imputation is carried out one variable at a time.

5.4.2 Direct Application

We applied our proposed methods as well as Lasso to the SOLVD-T trial, and provide a treatment covariate interaction recovery map in Figure 5.4. Each column represents a method while each row shows a candidate effect modifiers. A complete list of description of these variables is given in the Appendix. The black area indicates which candidates are included in each model, whereas the blank area

shows no treatment covariate interaction. It is clearly seen that $6, 5, 13, 7$ and 11 variables were predicted to influence the patient's response to treatment respectively by the four parameterization methods (5.4), (5.5), (5.6), and (5.7) and Lasso. Among our findings, left ventricular ejection fraction, "lvef", was found by all five methods to significantly modify the effect of enalapril on the survival outcome, with greater beneficial effect seen at the lower values of ejection fraction. Similar findings are also reported in the study of heterogeneous treatment response using the same trial data (Henderson et al., 2017). We are unable to assess the sensitivity / specificity of the methods in this real trial since the true effect modifiers are unknown, but it serves a good purpose to illustrate the use of our methodology for a real data application.

5.4.3 Extended Real Data-based Simulation

Despite the inability to evaluate the sensitivity / specificity on this real data, we instead engineered a design to assess a "partial specificity" of these methods. We create m noise variables that have no association at all with the response on top of the original data, and fit each method on this expanded dataset. This "partial specificity" is thus defined to be the percentage of the noise variables that are correctly identified by the method as having no interaction with the treatment, i.e., zero treatment covariate interaction. We set $m = 25$ and repeat the procedure 100 times. A summary of the average "partial specificity" for each method is provided

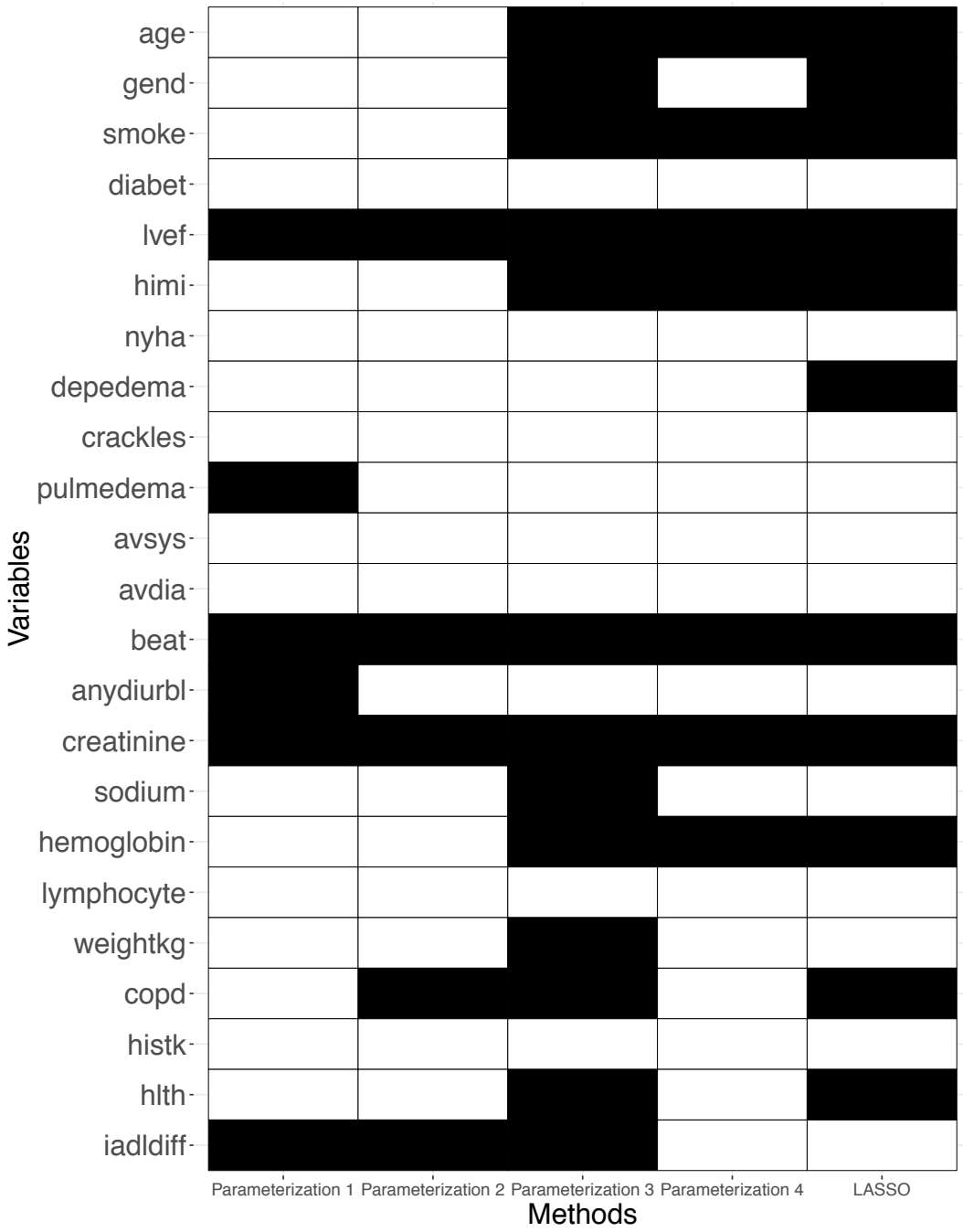


Figure 5.4: Treatment covariate interaction recovery map for the proposed methods and the Lasso.

in Table 5.4. An improvement regarding this partial specificity can be seen for all our proposed methods over Lasso. Lasso has the lowest value 62.6%, which says that for 100 added noise variables that are completely independent from the response, Lasso would falsely label around $100 - 63 = 37$ as the effect modifiers, predictive of the response. Surprisingly, Parameterization scheme 1 (5.4) and 2 (5.5) nearly reject all the noise variables, with close to 100% partial specificity. It is interesting to note that when the true data generating distribution is unknown, as in this real trial, Parameterization scheme 1 (5.4) and 2 (5.5) seem more robust to the model parametric assumptions than the other methods.

Table 5.4: The mean partial specificity of the identification of treatment covariate interactions for each method based on the SOLVD-T trial. P1, P2, P3 and P4 represents Parameterization 1, 2, 3 and 4 respectively.

	P1	P2	P3	P4	Lasso
Mean	99.2%	99.5%	65.8%	79.5%	62.6%

5.5 Remarks

Understanding how patients respond differently to treatment is key to translating clinical trial findings into decision makings for individuals regarding treatment recommendation. Prediction models for outcome prognosis are common in clinical research, however, models that predict treatment response heterogeneity in terms of individual baseline characteristics are much less common. Such mod-

els can be useful for answering the question of which treatment is better for a given individual. We provide a general prediction method to assess treatment response heterogeneity by adding novel, convex constraints to Lasso that honor the interaction hierarchy restriction. We give a clear interpretation on how individual characteristics affect treatment response by modeling pairwise interactions between treatment and covariates. We further extend the work of Kovalchik et al. (2013) by relaxing the constraint that the treatment covariate interactions are proportional to the main effects. Our proposed methods are able to automatically screen a larger number of candidate effect modifiers with the aid of parameter regularization inherited in Lasso, aiming to capture those with non-zero treatment covariate interactions as precisely as possible. The simulation study in Section 5.3 and the real data example shown in Section 5.4 demonstrate superior performance of our proposed methods against Lasso regarding the accuracy of risk prediction as well as the ability to correctly identify the non-zero interactions. In terms of choosing a parameterization scheme in practice, we recommend to use Parameterization scheme 1 (5.4) and 2 (5.5) as default choices because they exhibit the best performance in the real data application on SOLVD-T trial, where the true underlying data generating distribution remains unknown. However, in the case where the model parametric assumption can well approximate the true data generating distribution, e.g., as backed up by scientific theory, if there is a strong belief that the truth incorporates interaction hierarchy, all our proposed schemes are

desirable to use, though Parameterization scheme 4 (5.7) has a trade-off between unstable performance in risk prediction and the great ability to capture non-zero interactions (as shown by lowest GIRC). While in this case if the interaction hierarchy cannot be trusted, we suggest to use Parameterization scheme 4 (5.7) instead, which shows greater prediction accuracy and more precise identification of treatment effect modification over Lasso and the other proposed methods in scenario (B). Despite the focus of this chapter on TTE outcome, our methods can easily be adapted to other types of endpoints, for example, continuous or binary. All the users need to do is to change the objective function that needs to be optimized, for example, use instead the likelihood function for Gaussian distribution or Binomial distribution while the proposed constraints stay unchanged.

There are some limitations with applying our proposed methods for the analysis of treatment response heterogeneity and these limitations of course open room for future extension and improvement. First of all, the chapter focuses on the prediction side rather than inference, and that is because it is difficult to draw inferences on Lasso type of problem and very few approaches are available for this purpose. It would remain an area of future research to incorporate the inference, e.g., confidence intervals of the parameters, into the proposed prediction model with interaction hierarchy restriction. Nonetheless, our methods are well-suited for treatment covariate interaction selection purpose. Secondly, our proposed methods are built upon (semi)parametric model assumptions, such as the Cox model

(5.1). Vaughan et al. (2017) presents a stagewise estimation with generalized estimating equations to select interactions for the clustered data. The merit of using generalized estimating equations are twofold: (1) it does not require a parametric model assumption, and (2) it allows for the analysis of longitudinal/clustered data. It is possible to adapt our proposed methods with their stagewise procedure, thus relaxing the parametric model assumption, which is also likely to be more computationally efficient than the SPG method. Various methods exist to solve the constrained optimization problem, and one of the future work would be to find the ones with the highest computational efficiency. Since inference is not our focus, our proposed prediction methods lie in the category of exploratory analysis, and the significant effect modifiers found by the methods should be externally and independently validated in future studies.

Nonetheless, we suggest to incorporate our proposed methods into the secondary analyses of clinical trial, particularly Phase III trials, to make assessment of heterogeneity in patient's response to the experimental treatment. An R package is under development for implementation of the methods proposed here.

5.6 Appendix

5.6.1 Description of the Variables of the SOLVD-T Trial

Here we describe the 23 variables of the patients used as the candidate effect modifiers in the SOLVD-T trial. "age" is the patient's age at baseline, continuous; "gend" is patient's gender, binary, where 0 means female and 1 means male; "smoke" is a categorical variable, encoding patient's smoking history (0 = Never, 1 = Former and 2 = Current); "diabet" is an indicator of diabetes at baseline (1 = diabetes while 0 not); "lvef" is the baseline ejection fraction, continuous; "himi" is the history of myocardial infarction (1 = Yes, 0 = No); "nyha" is the New York Heart Association functional status at baseline, integer from 1 to 4 in the order of increasing risk; "depedema" is an indicator of whether or not the patient has dependent edema at baseline (1 = Yes, 0 = No); "crackles" is an indicator of whether or not the patient has crackles at baseline (1 = Yes, 0 = No); "pulmedema" is an indicator of whether or not the patient has pulmonary edema at baseline (1 = Yes, 0 = No); "avsys" is the baseline systolic blood pressure, continuous; "avdia" is the baseline diastolic blood pressure, continuous; "beat" is the baseline heartbeat, continuous; "anydiurbl" related to any use of a diuretic at baseline (1 = Yes, 0 = No); "creatinine", "hemoglobin", "lymphocyte" are three continuous clinical measures at baseline from the blood work; "weightkg" is the weight of the patient at baseline in kilograms; "copd" is an indicator of the chronic obstructive pulmonary disease

at baseline (1 = Yes, 0 = No); "histk" is history of any stroke at baseline (1 = Yes, 0 = No); "hlth" is self-rated health at baseline, integers from 1 to 5; "iadldiff" is any difficulty in performing instrumental activities of daily living, e.g., paying bills, grocery shopping, managing finances, where 1 = Yes, 0 = No.

Chapter 6

Discussion

In this dissertation, I developed four novel statistical methods and applications in Chapter 2 to 5 for a series of clinical trials, ranging from Phase I studies to Phase III studies that are used to determine whether a new experimental treatment/drug/device can go to the market or not. Chapter 2 presents a novel dose-finding design in Phase I studies, that jointly models an efficacy outcome and the toxicity data with multiple types over multiple treatment cycles. The extensive simulations conducted showed a high probability of finding the optimal doses and good overdose control. The design provides a relaxation of the traditional Phase I dose-finding methods that use only the binary toxicity data from the first treatment cycle and assume a monotone increasing relationship between dose and efficacy. It is very likely in practice that the experimental agent has an unknown relationship with dose, which should be explored during the study. As

a future work, I continue collaborating with Mayo Clinic on an extension of this work that incorporates Bayesian interval design and a novel construction of utility function that would drive the dose-finding algorithm.

Chapter 3 proposes and evaluates a two-stage, Phase II, adaptive clinical trial design. Its goal is to determine whether future Phase III (confirmatory) trials should be conducted, and if so, which population should be enrolled. The population selected for Phase III enrollment is defined in terms of a disease severity score measured at baseline. The design and analysis is optimized in a decision theory framework. The use of decision theory to guide development of an adaptive design (which I do here) is discussed in a guidance document on adaptive designs for medical devices by the Food and Drug Administration (FDA, 2010), who state "Adaptive designs that rely on anticipated regret can decrease the uncertainty in studies and make them much more predictable." One of the future works is to continue exploring the use of decision theory framework in improving the design.

In Chapter 4, I provide a general framework for adaptive enrichment designs in Phase III studies with delayed outcome, leveraging information in baseline variables and short-term outcomes to improve precision by using semiparametric, locally efficient estimators at each interim analysis. Through simulations of a real trial, I showed a substantial reduction in sample size needed/expected to conduct the trial, yet with comparable power, bias, variance and mean squared error against the standard estimator/design. It is an area of future work to apply

the advanced methods shown to have enhanced efficiency, e.g., the estimators of Lu and Tsiatis (2011); Rotnitzky et al. (2012); Gruber and van der Laan (2012), in the context of adaptive enrichment designs, leveraging baseline information and short-term outcomes.

A novel prediction method for treatment response heterogeneity is proposed in Chapter 5, as a secondary analysis in Phase III studies. The objective of the proposed methods is to select a subset of the pool of potential effect modifiers that have interactions with treatment and provide the estimates. The goal is achieved by adding convex constraints to Lasso that honor interaction hierarchy and a sparse coefficient structure. The simulation study in Section 5.3 and the real data example shown in Section 5.4 demonstrated superior performance of our proposed methods against Lasso regarding the accuracy of risk prediction as well as the ability to correctly identify the non-zero interactions. It is an area of future work to adapt our proposed methods with the stagewise procedure, as proposed in Vaughan et al. (2017), in order to relax the parametric model assumption.

Another interesting area to work on in future is the seamless trial design (e.g., Hampson and Jennison (2015)), which requires a complete protocol to be specified at the outset. This not only saves time in between trials, but also has a great potential to improve power of the trial, by recycling the trial participants information during the decision-making process. This potentially leads to huge savings in conducting the trial and a faster process to move a molecule to regulatory approval.

Bibliography

Alexander, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab. 12*(4), 1041–1067.

Bailey, K. R. (1994). Generalizing the results of randomized clinical trials. *Controlled Clinical Trials 15*(1), 15 – 23.

Banerjee, A. and A. A. Tsiatis (2006). Adaptive two-stage designs in phase ii clinical trials. *Statistics in Medicine 25*(19), 3382–3395.

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*(4), 962–973.

Barker, A., C. Sigman, G. Kelloff, N. Hylton, D. Berry, and L. Esserman (2009). I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy. *Clinical Pharmacology & Therapeutics 86*(1), 97–100.

Barzilai, J. and J. M. Borwein (1988). Two-point step size gradient methods. *IMA Journal of Numerical Analysis 8*(1), 141–148.

BIBLIOGRAPHY

Bauer, P. (1989). Multistage testing with adaptive designs (with discussion). *Biometrie und Informatik in Medizin und Biologie 20*, 130–148.

Bauer, P. and K. Köhne (1994). Evaluations of experiment s with adaptive interim analyses. *Biometrics 50*, 1029–1041.

Bekele, B. N. and Y. Shen (2005). A bayesian approach to jointly modeling toxicity and biomarker expression in a phase i/ii dose-finding trial. *Biometrics 61*(2), 344–354.

Bekele, B. N. and P. F. Thall (2004). Dose-finding based on multiple toxicities in a soft tissue sarcoma trial. *Journal of the American Statistical Association 99*(465), 26–35.

Bertsekas, D. (1976, Apr). On the goldstein-levitin-polyak gradient projection method. *IEEE Transactions on Automatic Control 21*(2), 174–184.

Bien, J., J. Taylor, and R. Tibshirani (2013, 06). A lasso for hierarchical interactions. *Ann. Statist. 41*(3), 1111–1141.

Birgin, E., J. Martínez, and M. Raydan (2014). Spectral projected gradient methods: Review and perspectives. *Journal of Statistical Software, Articles 60*(3), 1–21.

Birgin, E. G., J. M. Martínez, and M. Raydan (2000). Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization 10*(4), 1196–1211.

BIBLIOGRAPHY

Boessen, R., F. van der Baan, R. Groenwold, A. Egberts, O. Klungel, D. Grobbee, M. Knol, and K. Roes (2013). Optimizing trial design in pharmacogenetics research: comparing a fixed parallel group, group sequential, and adaptive selection design on sample size requirements. *Pharmaceutical Statistics*.

Brannath, W., E. Zuber, M. Branson, F. Bretz, P. Gallo, M. Posch, and A. Racine-Poon (2009). Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine 28*(10), 1445–1463.

Braun, T. M., P. F. Thall, H. Nguyen, and M. de Lima (2007). Simultaneously optimizing dose and schedule of a new cytotoxic agent. *Clinical Trials 4*(2), 113–124. PMID: 17456511.

Bretz, F., H. Schmidli, F. König, A. Racine, and W. Maurer (2006). Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal 48*(4), 623–634.

Brockwell, A. E. and J. B. Kadane (2003). A gridding method for bayesian sequential decision problems. *Journal of Computational and Graphical Statistics 12*(3), 566–584.

Cai, T., L. Tian, P. H. Wong, and L. J. Wei (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics 12*(2), 270–282.

Carlin, B. P., J. B. Kadane, and A. E. Gelfand (1998). Approaches for optimal sequential decision analysis in clinical trials. *Biometrics*, 964–975.

Cheng, Y. and D. A. Berry (2007). Optimal adaptive randomized designs for clinical trials. *Biometrika 94*(3), 673–689.

Chipman, H. (1996). Bayesian variable selection with related predictors. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique 24*(1), 17–36.

Colton, T. (1963). A model for selecting one of two medical treatments. *Journal of the American Statistical Association 58*(302), 388–400.

Colton, T. (1965). A two-stage model for selecting one of two treatments. *Biometrics 21*(1), 169–180.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 187–220.

Cox, D. R. (1984). Interaction. *International Statistical Review / Revue Internationale de Statistique 52*(1), 1–24.

Ding, M., G. L. Rosner, and P. Müller (2008). Bayesian optimal design for phase ii screening trials. *Biometrics 64*(3), 886–894.

Emerson, S. C., K. D. Rudser, and S. S. Emerson (2011). Exploring the benefits of adaptive sequential designs in time-to-event endpoint settings. *Statistics in medicine 30*(11), 1199–1217.

BIBLIOGRAPHY

Ezzalfani, M., S. Zohar, R. Qin, S. J. Mandrekar, and M.-C. L. Deley (2013). Dose-finding designs using a novel quasi-continuous endpoint for multiple toxicities. *Statistics in Medicine 32*(16), 2728–2746.

FDA (2010). Draft guidance for industry. Adaptive design clinical trials for drugs and biologics. https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf.

FDA (2015). Draft guidance for industry. adaptive designs for medical device clinical studies. http://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/ guidancedocuments/ucm446729.pdf.

Follmann, D. A. and M. A. Proschan (1999). A multivariate test of interaction for use in clinical trials. *Biometrics 55*(4), 1151–1155.

Freidlin, B., W. Jiang, and R. Simon (2010). The cross-validated adaptive signature design. *Clinical Cancer Research 16*(2), 691–698.

Freidlin, B. and R. Simon (2005). Adaptive signature design: An adaptive clinical-trial design for generating and prospectivelytesting agene expression signature for sensitive patients. *Clin Cancer Res 11*, 7872–7878.

Friede, T., N. Parsons, and N. Stallard (2012). A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in medicine 31*(30), 4309–4320.

Genz, A., F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn (2014). mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-0. URL http://CRAN.R-project.org/package=mvtnorm.

Götte, H., M. Donica, and G. Mordenti (2015). Improving probabilities of correct interim decision in population enrichment designs. *Journal of biopharmaceutical statistics 25*(5), 1020–1038.

Graf, A. C., M. Posch, and F. Koenig (2015). Adaptive designs for subpopulation analysis optimizing utility functions. *Biometrical Journal 57*(1), 76–89.

Gruber, S. and M. van der Laan (2012). Targeted minimum loss based estimator that outperforms a given estimator. *The International Journal of Biostatistics 8*(1).

Hampson, L. V. and C. Jennison (2013). Group sequential tests for delayed responses (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol. 75*(1), 3–54.

Hampson, L. V. and C. Jennison (2015). Optimizing the data combination rule for seamless phase ii/iii clinical trials. *Statistics in Medicine 34*(1), 39–58.

Hanley, D. (2012). http://braininjuryoutcomes.com/studies/mistie/entry/mistie/internatio stroke-conference-2012-mistie-phase-2-results.

Hanley, D. F., R. E. Thompson, J. Muschelli, M. Rosenblum, N. McBee, K. Lane, A. J. Bistran-Hall, S. W. Mayo, P. Keyl, D. Gandhi, T. C. Morgan, N. Ullman, W. A. Mould, J. R. Carhuapoma, C. Kase, W. Ziai, C. B. Thompson, G. Yenokyan,

E. Huang, W. C. Broaddus, R. S. Graham, E. F. Aldrich, R. Dodd, C. Wijman, J.-L. Caron, J. Huang, P. Camarata, D. Mendelow, B. Gregson, S. Janis, P. Vespa, N. Martin, I. Awad, and M. Zuccarello (2016). Safety and efficacy of minimally invasive surgery plus alteplase in intracerebral haemorrhage evacuation (MISTIE): a randomised, controlled, open-label, phase 2 trial. *The Lancet Neurology 15*(12), 1228–1237.

Harrell, F. E., K. L. Lee, and D. B. Mark (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine 15*(4), 361–387.

Henderson, N. C., T. A. Louis, G. L. Rosner, and R. Varadhan (2017, June). Individualized Treatment Effects with Censored Data via Fully Nonparametric Bayesian Accelerated Failure Time Models. *ArXiv e-prints*.

Henderson, N. C., T. A. Louis, C. Wang, and R. Varadhan (2016, September). Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research. *Health Services & Outcomes Research Methodology 16*(4), 213–233.

Houede, N., P. F. Thall, H. Nguyen, X. Paoletti, and A. Kramar (2010). Utility-based optimization of combination therapy using ordinal toxicity and efficacy in phase i/ii trials. *Biometrics 66*(2), 532–540.

Jenkins, M., A. Stone, and C. Jennison (2011). An adaptive seamless phase ii/iii

design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics 10*(4), 347–356.

Jennison, C. and B. W. Turnbull (1999). *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC Press.

Jennison, C. and B. W. Turnbull (2007). Adaptive seamless designs: Selection and prospective testing of hypotheses. *J. Biopharmaceutical Statistics*, 1135–1161, doi: 10.1080/10543400701645215.

Jiang, W., B. Freidlin, and R. Simon (2007). Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute 99*(13), 1036–1043.

Kent, D. and R. Hayward (2007). Limitations of applying summary results of clinical trials to individual patients: The need for risk stratification. *JAMA 298*(10), 1209–1212.

Kim, E. S., R. S. Herbst, I. I. Wistuba, J. J. Lee, G. R. Blumenschein, A. Tsao, D. J. Stewart, M. E. Hicks, J. Erasmus, S. Gupta, et al. (2011). The BATTLE trial: personalizing therapy for lung cancer. *Cancer discovery 1*(1), 44–53.

Kovalchik, S. A., R. Varadhan, and C. O. Weiss (2013). Assessing heterogeneity of treatment effect in a clinical trial with the proportional interactions model. *Statistics in Medicine 32*(28), 4906–4923.

Kraft, D. (1988). A software package for sequential quadratic programming. *Technical Report DFVLR-FB 88-28*.

Krisam, J. and M. Kieser (2015). Optimal decision rules for biomarker-based subgroup selection for a targeted therapy in oncology. *International Journal of Molecular Sciences 16*(5), 10354.

Lai, T. L., P. W. Lavori, and O. Y.-W. Liao (2014). Adaptive choice of patient subgroup for comparing two treatments. *Contemporary Clinical Trials 39*(2), 191 – 200.

Lan, K. K. G. and D. L. DeMets (1983). Discrete sequential boundaries for clinical trials. *Biometrika 70*, 659–663.

Lee, J. J., X. Gu, and S. Liu (2010). Bayesian adaptive randomization designs for targeted agent development. *Clinical Trials 7*(5), 584–596.

Lee, S. M., D. L. Hershman, P. Martin, J. P. Leonard, and Y. K. Cheung (2012). Toxicity burden score: a novel approach to summarize multiple toxic effects. *Annals of Oncology 23*(2), 537–541.

Legedza, A. T. and J. G. Ibrahim (2000). Longitudinal design for phase i clinical trials using the continual reassessment method. *Controlled Clinical Trials 21*(6), 574 – 588.

Lehmacher, W. and G. Wassmer (1999). Adaptive sample size calculations in group sequential trials. *Biometrics 55*(4), 1286–1290.

Liu, Q., M. A. Proschan, and G. W. Pledger (2002). A unified theory of two-stage adaptive designs. *JASA 97*(460), 1034–1041.

Lu, X. and A. A. Tsiatis (2011). Semiparametric estimation of treatment effect with time-lagged response in the presence of informative censoring. *Lifetime data analysis 17*(4), 566–593.

Morgan, T., M. Zuccarello, R. Narayan, P. Keyl, K. Lane, and D. Hanley (2008). Preliminary findings of the minimally-invasive surgery plus rtpa for intracerebral hemorrhage evacuation (mistie) clinical trial. *Acta Neurochir Suppl. 105*, 147–51.

Müller, H.-H. and H. Schäfer (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics 57*(3), 886–891.

Nelder, J. A. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General) 140*(1), 48–77.

Ohwada, S. and S. Morita (2016). Bayesian adaptive patient enrollment restriction to identify a sensitive subpopulation using a continuous biomarker in a randomized phase 2 trial. *Pharmaceutical Statistics*.

Parast, L., L. Tian, and T. Cai (2014). Landmark estimation of survival and treatment effect in a randomized clinical trial. *J. Amer. Statist. Assoc. 109*(505), 384–394.

PCORI (2013). The PCORI (Patient-Centered Outcomes Research Institute) Methodology Report pcori.org/research-we-support/research-methodology-standards.

Peixoto, J. L. (1987). Hierarchical variable selection in polynomial regression models. *The American Statistician 41*(4), 311–313.

Pencina, M. J. and R. B. D'Agostino (2004). Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine 23*(13), 2109–2123.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raydan, M. (1997). The barzilai and borwein gradient method for the large scale unconstrained minimization problem. *SIAM Journal on Optimization 7*(1), 26–33.

Robins, J. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*.

Robins, J. M. and A. Rotnitzky (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS Epidemiology-Methodological Issues 297–331*.

Rosenblum, M., B. Luber, R. E. Thompson, and D. Hanley (2016). Group sequential designs with prospectively planned rules for subpopulation enrichment. *Statistics in Medicine 35*(21), 3776–3791. sim.6957.

Rosenblum, M., T. Qian, Y. Du, H. Qiu, and F. Aaron (2016). Multiple testing procedures for adaptive enrichment designs: combining group sequential and reallocation approaches. *Biostatistics 17*(4), 650–662.

Rossell, D., P. Müller, and G. L. Rosner (2006). Screening designs for drug development. *Biostatistics 8*(3), 595–608.

Rothwell, P. M. (1995, 2018/01/26). Can overall results of clinical trials be applied to all patients? *The Lancet 345*(8965), 1616–1619.

Rothwell, P. M. (2005, 2018/01/26). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet 365*(9454), 176–186.

Rotnitzky, A., Q. Lei, M. Sued, and J. Robins (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika 99*(2), 439–456.

Rubinstein, L. V., E. L. Korn, B. Freidlin, S. Hunsberger, S. P. Ivy, and M. A. Smith (2005). Design issues of randomized phase ii trials and a proposal for phase ii screening trials. *Journal of Clinical Oncology 23*(28), 7199–7206. PMID: 16192604.

Scharfstein, D., A. Tsiatis, and J. Robins (1997). Semiparametric efficiency and its implications on the design and analysis of group-sequential studies. *Journal of the American Statistical Association 92*(440), 1342–1350.

Schmidli, H., F. Bretz, A. Racine, and W. Maurer (2006). Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: applications and practical considerations. *Biometrical Journal 48*(4), 635–643.

Schwab, J., S. Lendle, M. Petersen, and M. van der Laan (2014). *ltmle: Longitudinal Targeted Maximum Likelihood Estimation*. R package version 0.9.3-1.

Slud, E. V. and L.-J. Wei (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J. Amer. Statist. Assoc. 77*, 862–868.

Spencer, A. V., C. Harbron, A. Mander, J. Wason, and I. Peers (2016). An adaptive design for updating the threshold value of a continuous biomarker. *Statistics in Medicine*.

Stallard, N., T. Hamborg, N. Parsons, and T. Friede (2014). Adaptive designs for confirmatory clinical trials with subgroup selection. *Journal of biopharmaceutical statistics 24*(1), 168–187.

Thall, P. F. and J. D. Cook (2004). Dose-finding based on efficacy–toxicity trade-offs. *Biometrics 60*(3), 684–693.

The-SOLVD-Investigators (1991). Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *New England Journal of Medicine 325*(5), 293–302. PMID: 2057034.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

van der Laan, M. and S. Rose (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Berlin Heidelberg New York: Springer.

van der Laan, M. and D. Rubin (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics 2*(1).

van der Laan, M. J. and S. Gruber (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics 8*(9).

Varadhan, R. and P. Gilbert (2009). Bb: An r package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software, Articles 32*(4), 1–26.

Varadhan, R. and C. Roland (2008). Simple and globally convergent methods for

accelerating the convergence of any em algorithm. *Scandinavian Journal of Statistics 35*(2), 335–353.

Vaughan, G., R. Aseltine, K. Chen, and J. Yan (2017). Stagewise generalized estimating equations with grouped variables. *Biometrics 73*(4), 1332–1342.

Wang, C. Y., N. Wang, and S. Wang (2000). Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. *Biometrics 56*(2), 487–495.

Wang, S. J., H. Hung, and R. T. O'Neill (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal 51*, 358–374.

Wang, S. J., R. T. O'Neill, and H. Hung (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subsets. *Pharmaceut. Statist. 6*, 227–244.

Xu, Y., L. Trippa, P. Müller, and Y. Ji (2014). Subgroup-based adaptive (SUBA) designs for multi-arm biomarker trials. *Statistics in Biosciences*, 1–22.

Yin, J., R. Qin, M. Ezzalfani, D. J. Sargent, and S. J. Mandrekar (2017). A bayesian dose-finding design incorporating toxicity data from multiple treatment cycles. *Statistics in Medicine 36*(1), 67–80. sim.7134.

Zhang, M. (2015). Robust methods to improve efficiency and reduce bias in esti-

mating survival curves in randomized clinical trials. *Lifetime Data Analysis 21*(1), 119–137.

Zhou, X., S. Liu, E. S. Kim, R. S. Herbst, and J. J. Lee (2008). Bayesian adaptive design for targeted therapy development in lung cancer–a step toward personalized medicine. *Clinical Trials 5*(3), 181–193.

Vita

Yu Du was born on May 5th, 1988, in Beijing, China. He acquired his Bachelor of Arts degree in Management Information System from Beijing University of Technology in China. During his bachelor program, he went to study at Australian National University in 2009 for a year as an exchange student. Upon graduation, he joined the two-year master program in Department of Applied Mathematics and Statistics at Johns Hopkins University. He acquired his Master of Science degree in Engineering in 2013 and began his PhD degree journey in the fall of 2013 at Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health. His research interests focus primarily on clinical trial designs, covering the development and application of novel statistical methods for Phase I studies, Phase II studies and Phase III studies. He is also interested and did research in algorithm acceleration as well as wearable computing, with applications in the clinical setting. He was the recipient of departmental First-year PhD Examination Award in 2014, June B. Culley Award in 2016 that honors outstanding achievement on preliminary school-wide examination paper and Margaret Merrell Award in 2018, the

department highest recognition for outstanding research by a Biostatistics doctoral student. He was named United States Food and Drug Administration (FDA) and Johns Hopkins Center of Excellence in Regulatory Science and Innovation (CERSI) Scholar from 2015 to 2018. He got married to Shuyuan Wu in Baltimore, 2013 and Will(iam) Muyi Du, his son, was born also in Baltimore on December 8th, 2017.

VITA

EDUCATION AND TRAINING

2018 **Johns Hopkins University**, Baltimore, MD
Ph.D. in Biostatistics
Thesis title: *The Statistical Methods and Applications in Clinical Trials*
Advisors: Michael Rosenblum, Ravi Varadhan *and* Vadim Zipunnikov

2013 **Johns Hopkins University**, Baltimore, MD
M.S. in Applied Mathematics and Statistics

2011 **Beijing University of Technology**, Beijing, China
B.A. in Management Information System

Australian National University, Canberra, Australia
Non-degree Exchange Program

PUBLICATIONS AND MANUSCRIPTS

PUBLISHED

1. Segal, J., Chang, H., **Du, Y.**, Walston, J., Carlson, M., *and* Varadhan, R. (2017) Development of a Claims-Based Frailty Indicator Using a Well-Established Frailty Phenotype. *Medical Care*, 55(7), 716–722.

2. Seyednasrollah F et al. - **Du, Y.** in PCC DREAM Consortium. (2017) A DREAM Challenge to Build Prediction Models for Short-Term Discontinuation of Docetaxel in Metastatic Castration-Resistant Prostate Cancer. *JCO Clinical Cancer Informatics*, 2017:1, 1-15

3. Rosenblum, M., Qian, T., **Du, Y.**, Qiu, H., *and* Fisher, A. (2016) Multiple Testing Procedures for Adaptive Enrichment Designs: Combining Group Sequential and Reallocation Approaches. *Biostatistics*, 17(4), 650–662.

4. Deng, D., **Du, Y.**, Ji, Z., Rao, K., Wu, Z., Zhu, Y., *and* Coley, R. (2016) Predicting Survival Time for Metastatic Castration Resistant Prostate Cancer: An Iterative Imputation Approach. *F1000Research*, 5:2672 (doi: 10.12688/f1000research.8628.1)

5. Guinney J et al. - **Du, Y.** in PCC DREAM Consortium (2016). Prediction of Overall Survival for Patients with Metastatic Castration-resistant Prostate Cancer: Development of a Prognostic Model through a Crowdsourced Challenge with Open Clinical Trial Data. *The Lancet Oncology*, 18(1), 132–142

UNDER REVIEW / IN PREPARATION

6. **Du, Y.**, Jun, Y., Sargent, D., *and* Mandrekar, S. An Adaptive Multi-Stage Phase I Dose-finding Design Incorporating Continuous Efficacy and Toxicity Data from Multiple Treatment Cycles. Under review at *Journal of Biopharmaceutical Statistics* (revision requested December 15th, 2017).

7. Yin, J., **Du, Y.**, Sargent, D., Mandrekar, S., *and* Qin, R. phase1RMD: an R Package for Repeated Measures Dose-finding Designs with Novel Toxicity and Efficacy Endpoints. Under review at *Computer Methods and Programs in Biomedicine*.

8. **Du, Y.**, *and* Varadhan, R. SQUAREM : An R Package for Off-the-Shelf Acceleration of EM, MM and other EM-like Monotone Algorithms. Under Review at *R Journal*.

9. **Du, Y.**, Rosner, G., *and* Rosenblum, M. Phase II Adaptive Enrichment Design to Determine the Population to Enroll in Phase III Trials, by Selecting Thresholds for Baseline Disease Severity. *Johns Hopkins University, Dept. of Biostatistics Working Papers.* Working Paper 290. (http://biostats.bepress.com/jhubiostat/paper290)

10. **Du, Y.**, Qian, T., *and* Rosenblum, M. Bias, Variance, and Sample Size Reductions due to Adjustment for Prognostic Baseline Variables and Short Term Outcomes in Adaptive Enrichment Trial Designs with Delayed Outcomes. *Working Paper*.

11. **Du, Y.**, *and* Varadhan, R. Lasso Estimation of Hierarchical Interactions for Analyzing Heterogeneous Treatment Effect. *Working Paper*.

12. **Du, Y.**, Lin, Y., Wang, Y., *and* Chiang, A. Adaptive Seamless Phase I/II/III Trial Design in Oncology. *Working Paper*.

13. Zipunnikov, V., Li, X., **Du, Y.**, Zhu, Y., Green, P., Harris, T., *and* Maurer, M. Physical Disability Score for Real-time Monitoring of Physical Activity in Patients with Congestive Heart Failure. *Working Paper*.

VITA

PRESENTATIONS

2018 Lasso Estimation of Hierarchical Interactions for Analyzing Heterogeneous Treatment Effect. *ENAR*, Atlanta, GA

2017 Validation of Prediction Models: a Case Study. *ASA Biopharmaceutical Section Statistics Workshop*, Washington DC

2017 Phase II Adaptive Enrichment Design to Select the Population for Phase III Trials, in Terms of Baseline Disease Severity. *Joint Conference on Biometrics & Biopharmaceutical Statistics*, Vienna, Austria

2017 Adaptive Seamless Trial Designs in Oncology. *Eli Lilly Oncology Seminar*, Indianapolis, IN

2017 An Adaptive Multi-Stage Phase I Dose-finding Design Incorporating Continuous Efficacy and Toxicity Data from Multiple Treatment Cycles. *ENAR*, Washington DC

2017 An Adaptive Multi-Stage Phase I Dose-finding Design Incorporating Continuous Efficacy and Toxicity Data from Multiple Treatment Cycles. *The Informal Biostatistics Meeting*, Johns Hopkins School of Medicine, Baltimore, MD

2016 An Adaptive Multi-Stage Phase I Dose-finding Design Incorporating Continuous Efficacy and Toxicity Data from Multiple Treatment Cycles. *International Chinese Statistical Association*, Shanghai, China

2016* Deriving a Claims-Based Frailty Indicator with the Gold Standard Measure. *Annual Research Meeting of the Academic Health*, Boston, MA

2016* Towards Generating a Claims-based Frailty Indicator. *Annual Scientific Meeting of the American Geriatrics Society*, Long Beach, CA

2015 Optimal Two-Stage Adaptive Enrichment Design for Patient Population Recommendation Based on an Ordinal Risk Score in Planning Phase. *Individualized Health Initiative(InHealth) Methodology Group Meeting*, Baltimore, MD

2015 Optimal Two-Stage Adaptive Enrichment Design for Patient Population Recommendation Based on an Ordinal Risk Score in Planning Phase. *Joint Statistical Meeting*, Seattle, WA

VITA

AWARDS AND HONORS

1. Margaret Merrell Award (the department highest recognition for outstanding research by a Biostatistics doctoral student), 2018
2. United States Food and Drug Administration (FDA) and Johns Hopkins Center of Excellence in Regulatory Science and Innovation (CERSI) Scholar, 2018
3. United States Food and Drug Administration (FDA) and Johns Hopkins Center of Excellence in Regulatory Science and Innovation (CERSI) Scholar, 2017
4. June B. Culley Award (honors outstanding achievement on preliminary school-wide examination paper), 2016
5. United States Food and Drug Administration (FDA) and Johns Hopkins Center of Excellence in Regulatory Science and Innovation (CERSI) Scholar, 2016
6. Howard Hughes Medical Institute 2016 International Student Fellowship Nominee by Johns Hopkins Internal Review Committee (Only 7 are nominated), 2016
7. Best Performing Team in Prostate Cancer Dream Challenge 1b, 2015
8. First-year PhD Examination Award (the top performer), 2014
9. Professor Joel Dean Award for Excellence in Teaching, 2013

PROFESSIONAL ACTIVITIES

1. Associate Editor for *Biostatistics and Biometrics Open Access Journal (BBOAJ)*
2. Referee for *Statistics in Biosciences*; *Journal of General Internal Medicine*
3. Facilitator of JHU Data Science Hackathon (DaSH), 2015
4. Organizer of Computing Club, at Department of Biostatistics, JHU, 2014

TEACHING EXPERIENCE

2018 140.624 (Statistical Methods in Public Health IV). Johns Hopkins Bloomberg School of Public Health.

* indicates co-author and not the primary presenter

VITA

2017 140.622 (Statistical Methods in Public Health II). Johns Hopkins Bloomberg School of Public Health.

2017 140.621 (Statistical Methods in Public Health I). Johns Hopkins Bloomberg School of Public Health. *Lead TA*

2017 140.623 (Statistical Methods in Public Health III). Johns Hopkins Bloomberg School of Public Health. *Lead TA*

2016 140.622 (Statistical Methods in Public Health II). Johns Hopkins Bloomberg School of Public Health. *Lead TA*

2016 140.624 (Statistical Methods in Public Health IV). Johns Hopkins Bloomberg School of Public Health.

2016 140.624 (Statistical Methods in Public Health III). Johns Hopkins Bloomberg School of Public Health.

2015 140.624 (Statistical Methods in Public Health II). Johns Hopkins Bloomberg School of Public Health.

2015 140.624 (Statistical Methods in Public Health I). Johns Hopkins Bloomberg School of Public Health.

2015 140.624 (Statistical Methods in Public Health IV). Johns Hopkins Bloomberg School of Public Health.

2015 140.624 (Statistical Methods in Public Health III). Johns Hopkins Bloomberg School of Public Health.

2014 280.345 (Public Health Biostatistics). Johns Hopkins Bloomberg School of Public Health.

2013 550.445 (Financial Derivatives II). Department of Applied Mathematics and Statistics, Johns Hopkins University.

2012 550.444 (Financial Derivatives I). Department of Applied Mathematics and Statistics, Johns Hopkins University.